UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE SÃO JOÃO DA BOA VISTA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LUCAS CARDINAL DANTAS

Multi-objective optimization of a neural network-based nonlinear equalizer in unrepeated digital coherent optical communication systems

São João da Boa Vista
2024

Programa de Pós-Graduação Interunidades em Engenharia Elétrica

LUCAS CARDINAL DANTAS

Multi-objective optimization of a neural network-based nonlinear equalizer in unrepeated digital coherent optical communication systems

Dissertação apresentada ao Conselho de Pós-graduação em Engenharia Elétrica do Campus de São João da Boa Vista, Universidade Estatual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de mestre em Pós-graduação em Engenharia Elétrica .

Linha de Pesquisa: Sinais e Comunicação

Orientador: Prof. Dr. Ivan Aritz Aldaya Garde
Coorientador: Prof. Dr. Hildo Guillardi Júnior

São João da Boa Vista
2024

D192m

Dantas, Lucas Cardinal
Multi-objective optimization of a neural network-based nonlinear equalizer in unrepeated digital coherent optical communication systems / Lucas Cardinal Dantas. -- São João da Boa Vista, 2024
62 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Faculdade de Engenharia, São João da Boa Vista
Orientador: Ivan Aritz Aldaya Garde
Coorientador: Hildo Guillardi Júnior

1. Comunicações óticas. 2. Inteligência artificial. 3. Kerr, Efeito de. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Potential impact of this research

This work addresses an important problem in optical communication systems: the nonlinear distortion introduced by the fiber in coherent optical systems. Machine learning techniques were applied to process adjacent symbols instead of only the symbol of interest. Additionally, this research studied chromatic dispersion and concepts of signal modulation, such as, 16-QAM signals.

Impacto potencial desta pesquisa

Este trabalho aborda um problema importante em sistemas de comunicação óptica: a distorção não linear introduzida pela fibra em sistemas ópticos coerentes. Técnicas de aprendizado de máquina foram aplicadas para processar símbolos adjacentes ao invés de apenas o símbolo de interesse. Adicionalmente, esta pesquisa estudou a dispersão cromática e conceitos de modulação de sinais, por exemplo, sinais 16-QAM.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de São João da Boa Vista

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: Multi-objective Optimization of a Neural Network-based Nonlinear Equalizer in Unrepeated Digital Coherent Optical Communication Systems

AUTOR: LUCAS CARDINAL DANTAS
ORIENTADOR: IVAN ARITZ ALDAYA GARDE
COORIENTADOR: HILDO GUILLARDI JUNIOR

Aprovado como parte das exigências para obtenção do Título de Mestre em Engenharia Elétrica, área: Sistemas Eletrônicos pela Comissão Examinadora:

Prof. Dr. IVAN ARITZ ALDAYA GARDE (Participaçao Virtual)
Departamento de Engenharia Eletronica e de Telecomunicacoes / Faculdade de Engenharia de Sao Joao da Boa Vista UNESP

Profa. Dra. GRETHELL GEORGINA PÉREZ SÁNCHEZ (Participaçao Virtual)
División de Ciencias Básicas e Ingeniería / Universidad Autónoma Metropolitana (México)

Prof. Dr. PLINIO SANTINI DESTER (Participaçao Virtual)
Departamento de Engenharia Eletrônica e de Telecomunicações / Câmpus de São João da Boa Vista

São João da Boa Vista, 28 de novembro de 2024

Verônica Liberali Messias
Supervisora Técnica de Seção
Seção Técnica de Graduação e Pós-Graduação

Assinado de forma digital por Veronica Liberali Messias:36823715839
Dados: 2024.12.04 08:42:08 -03'00'

ACKNOWLEDGEMENTS

I would like to thank God for guiding my steps to this point through a long set of coincidences in my life. I also thank my parents, Luis and Marcia, and my brother Matheus for always providing me with the resources and continuously encouraging me to study more and more, without ever asking for anything in return. Without this encouragement, my journey in the study of engineering would not have even begun. I also thank Wallesca for all the love and understanding throughout this year.

I extend my gratitude to my advisor, Prof. Dr. Ivan Aldaya Garde for the valuable advice, clarity, and elegance in teaching. A special thanks goes to Prof. Dr. Hildo Guillardi Jr for his time dedicated to guiding the development of this work and his patience.

To all my friends who encouraged me to continue progressing in my studies and helped make my life lighter along the way. Also I thank my engineering colleagues who motivate me directly or indirectly to study and learn more everyday.

Finally, I thank Unesp for providing a pleasant and enriching environment for its students. Many of its graduates now hold brilliant positions in academia or industry. I am truly grateful to this great institution and to all its professors and staff. Thank you so much!

"Won't that be grand? Computers and the programs will start thinking and the people will stop."

Tron 1982

ABSTRACT

In this work, we propose using neural networks to mitigate intersymbol interference introduced by the fiber in optical communication systems. This mitigation method is well known, nevertheless the majority of studies that we traced back in the literature process only the symbol of interest, and consequently, cannot compensate the intersymbol interference. The approach presented in this work includes also using adjacent symbols, as there is a correlation between these symbols distortion. But, this increases the computational complexity. For that reason, we also analyzed the computational complexity to train the neural network. We were not able to trace back in the literature works that performed this multi-objective analysis, taking into account the adjacent symbols for training the neural network and also the computational complexity.

KEYWORDS: optical communication; artificial intelligence; kerr effect.

RESUMO

Neste projeto, propomos a utilização de redes neurais para mitigação da interferência intersimbólica introduzida pela fibra em sistemas de comunicações ópticas coerentes digitais. Este método de mitigação é conhecido, no entanto a maioria dos estudos encontrados na literatura processam apenas o símbolo de interesse e consequentemente não conseguem compensar a interferência intersimbólica. A proposta apresentada neste trabalho envolve utilizar também os símbolos adjacentes, uma vez que existe correlação entre a distorção destes símbolos. No entanto isto aumenta a complexidade computacional. Por esta razão, analisamos também a complexidade computacional para efetuar o treinamento da rede. Não encontramos anteriormente na literatura trabalhos que efetuaram esta análise multi-objetiva, levando em conta os símbolos adjacentes no treinamento da rede neural e também a complexidade computacional.

PALAVRAS-CHAVE: comunicações óticas; inteligência artificial; kerr, efeito de.

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ASIC	Application-Specific Integrated Circuit
BER	Bit Error Ratio
XPM	Cross Phase Modulation
DWDM	Dense Wavelength-Division Multiplexing
DBP	Digital Back-Propagation
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
FLOP	Floating-Point Operation
FWM	Four Wave Mixing
ISI	Intersymbol Interference
LOP	Launched Optical Power
MLP	Multilayer Perceptron
m-PSK	m-ary Phase Shift Keying
PMD	Polarization Mode Dispersion
QAM	Quadrature Amplitude Modulation
ReLU	Rectified Linear Unit
SPM	Self Phase Modulation
SNR	Signal to Noise Ratio
WDM	Wavelength-Division Multiplexing

LIST OF SYMBOLS

$g(u)$	Activation function
θ	Bias term
A_k	Constellation point complex amplitude
k	Constellation point
L_{eff}	Effective interaction length
$d_j(k)$	Expected value during training
L	Fiber length
$h_f(z,t)$	Fiber linear impulse response
β_1	Group velocity term
a	Activation function numerator parameter
η	Learning rate
δ	Local gradient
α	Loss coefficient
x	NN input
l	NN layer
I_j^l	NN layer input
Y_i^l	NN layer output
$E(k)$	NN loss function
y	NN output
w_i	NN weight
γ	Nonlinear coefficient
ϕ_{NL}	Nonlinear phase shift
ϕ_{NL}^j	Nonlinear phase shift considering multiple channels
N_{input}	Number of neurons in the input layer
N_{output}	Number of neurons in the output layer

$N_{neurons}$	Number of neurons in the hidden layer
N	Number of symbols
N_{op1}	Number of operations in the input-hidden connection
N_{op2}	Number of operations in the output-hidden connection
$A(z,t)$	Optical amplitude
$P(z)$	Optical power as function of position z
P_j	Optical power for channel j experiencing SPM
P_M	Optical power for channel m contributing to XPM
Δ	Phase mismatch in FWM
β	Propagation constant
f_p	Pulse shape
n	Refractive index
w_j	Resonant frequency
β_2	Second order dispersion term
b	Activation function denominator parameter
c	Speed of light in vacuum
T_S	Symbol period
B_j	Strength of j-th ressonance
P_{in}	Input power
W_{ij}^l	Weight matrices

TABLE OF CONTENTS

1 INTRODUCTION

In this section we are going to present the motivation for this work, starting with the historical perspective since early optical communication systems and then discussing the actual technological scenario. After that, we present the contribution and the work structure.

1.1 Optical communication systems

The use of light for communication is not new: it dates back to antiquity, when many civilizations used mirrors, smoke signals and fire beacons to send messages. In 1792, French engineer Claude Chappe proposed using mechanically coded messages over distances around 100 km. This type of "optical telegraph" (an optical telegraph is a line of stations used for communicating textual information visually) was first implemented in July 1794 between Paris and Lille, and, by 1830 had expanded throughout Europe. The word "Telegraph" is derived from the Greek words tele (τῆλε) "distant" and graphein (γράφειν) "writing", and this optical telegraph was sometimes called a *Napoleonic semaphore* (SCHOFIELD, 2013). This early optical system was followed by electrical telegraph systems, given the rise of the electrical communication era, after 1830. With it, new coding techniques were developed, such as the nowadays popular Morse code. The invention of the telephone in 1876 by Alexander Graham Bell contributed to many advancements in electrical communication systems. Later, the use of coaxial cables significantly increased system capacity; the first coaxial cable system in 1940 was capable of transmitting 300 voice channels within a single television channel. Subsequently, microwave communication systems were implemented, utilizing electromagnetic carrier waves with frequencies between 1 and 10 GHz.

After the invention of the *Light Amplification by Stimulated Emission of Radiation* (laser) in 1960, researchers focused their attention on using lasers for optical communication. The laser provided a coherent light source, along with the possibility of modulation at high frequencies (SENIOR; JAMRO, 2009). In 1961, Elias Snitzer published a theoretical description of single-mode fibers for information transmission (SNITZER, 1961). In 1966, it was suggested that optical fibers were the most suitable medium for this type of communication. Concepts for optical communication via dielectric waveguides or optical fibers made from glass to prevent optical signal degradation caused by atmospheric interference were put forward around the same time period. Notably, these proposals were made by Kao and Hockham (KAO; HOCKHAM, 1966) and Werts (WERTS, 1966).

The progress made from 1975 to 2000 can be grouped in generations. The first-generation of lightwave systems operated close to 0.8 μm and used Gallium Arsenide ($GaAs$) semiconductor lasers. Later, in the 1970s it was realized that operating the systems in the wavelength region close to 1.3 μm, repeater spacing could be increased. This resulted in a worldwide effort to develop Indium Gallium

Arsenide Phosphide ($InGaAsP$) semiconductor lasers operating close to 1.3 μm, and that is what is called the second-generation of fiber-optic communication systems. The repeater spacing of the second-generation was limited by the fiber losses when operating at 1.3 μm. These losses could be minimal by either operating near 1.55 μm, or by limiting the laser spectrum to a single longitudinal mode. During the 1980s these two approaches were used, and by 1985 experiments indicated the possibility of transmitting information at bit rates up to 4 Gb/s over more than 100 km. Then, the third-generation systems became available commercially in 1990. A few years later, the fourth-generation of lightwave systems started, using optical amplification to increase the repeater spacing and Wavelength-Division Multiplexing (WDM) (see Appendix C). Finally, fifth-generation of fiber-optic communication systems focused in extending the wavelength range in which WDM systems would operate, increasing the capacity of those systems (AGRAWAL, 2000).

Expansion in communication capacity is essential. At times, this need becomes even more pronounced, as seen during the pandemic years. Telecommunication companies reported a significant increase in the amount of data required for long-haul optical communication channels due to the COVID-19 quarantine, with activities such as video conferencing, remote work, and studying from home (FELDMANN *et al.*, 2020). In light of this growing demand for higher capacity, increasing fiber transmission length without the need for repeaters, along with overall capacity enhancement, is an important goal for telecommunication engineers and researchers.

Coherent receivers (a receiver is considered coherent if it can recover both the amplitude and phase of the signal) began to be studied in the 1980s (OKOSHI, 1987). Superior to direct detection in terms of sensitivity, they initially seemed promising for increasing the transmission length of optical links without the need for signal regenerators. These systems appeared to offer higher performance; but, after all, they were considered impractical at the time due to their complexity and high cost. Not only that, they were also highly vulnerable to phase noise and polarization rotation (AGRAWAL, 2000). The interest in coherent optical communication systems was revived mainly due to the replacement of analog electronic and optoelectronic modules in coherent optical receivers with less expensive, higher-speed Application-Specific Integrated Circuits (ASICs) (ROUDAS, 2011). This advancement enabled the equalization of linear transmission impairments, such as Polarization Mode Dispersion (PMD) and chromatic dispersion. About twenty years later, coherent systems resurfaced in research. Coherent receivers were then capable of recovering the carrier phase in the digital domain, allowing for the detection of both the phase and amplitude of the optical signal. With coherent receivers, it became possible to adopt modulation formats with higher spectral efficiency, such as m-ary Phase Shift Keying (m-PSK) and Quadrature Amplitude Modulation (QAM). Furthermore, the interferometric nature of coherent receivers opened the possibility of exploiting polarization diversity (MELLO; BARBOSA, 2021).

Naturally, communication systems are affected by noise (see Appendix E). Thermal noise is stochastic and has a small coherence time, making it difficult to compensate for. While the effects of chromatic dispersion, PMD, misalignments in polarization, phase, and time are mitigated using Digital Signal Processing (DSP), the performance of digital coherent systems is ultimately limited by a combination of additive noise and nonlinear effects. Additive noise can be generated in the optical domain by optical amplifiers or in the electrical domain by the receiver photodetectors (COSTA, 2022).

Compensation of nonlinearity is crucial for increasing the capacity of optical communication channels. Then, it is not surprising that many nonlinear compensation techniques have been explored in recent years. As potential solutions to nonlinear impairments, researchers have proposed Digital Back-Propagation (DBP) algorithms (RAFIQUE *et al.*, 2011) and Volterra series nonlinear equalizers (GAO *et al.*, 2009), both of which involve numerous Fast Fourier Transform (FFT) operations (COOLEY; TUKEY, 1965). As network distances increase, the complexity of these techniques must also rise to achieve effective compensation performance. And another thing: accurately obtaining link parameters remains challenging. To overcome these difficulties, various machine learning-based alternatives have been recently proposed, ranging from k-means (COSTA, 2022) to neural networks (ZHANG *et al.*, 2019). However, most of these techniques do not account for Intersymbol Interference (ISI). In this work, we propose an approach that combines adjacent symbols with machine learning techniques to address this issue.

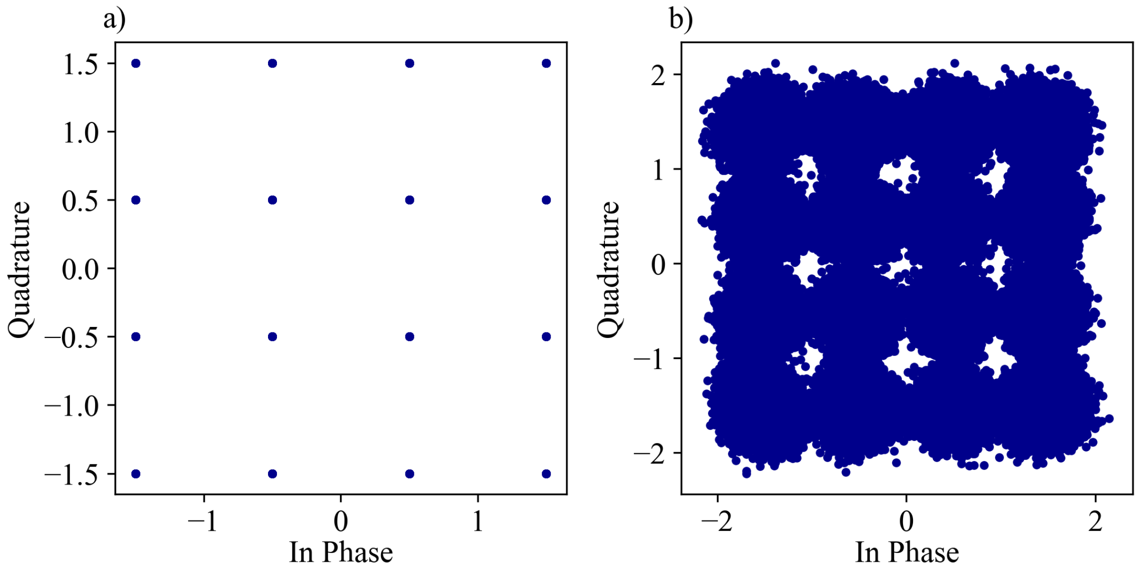
The nonlinear impairment between consecutive symbols is correlated (LUO *et al.*, 2022). Therefore, accounting for the nonlinear influence of both preceding and succeeding adjacent symbols on the current symbol may enhance the neural network's ability to capture complex patterns. Currently, most machine learning techniques process information exclusively based on the symbol of interest, neglecting both preceding and succeeding symbols, which may limit their capacity for effective mitigation. In this work, we aim to leverage this correlation to our advantage. In summary, we expect that by utilizing the correlation of nonlinear impairments between preceding and succeeding adjacent symbols, we can develop more accurate nonlinear optical equalizers.

The fiber is a nonlinear medium due to the Kerr effect, meaning that the refractive index depends on the intensity of the transmitted signal. This leads to phenomena such as Self-Phase Modulation (SPM), Cross-Phase Modulation (XPM), Four-Wave Mixing (FWM), and modulation instability (LI; LI; NOLAN, 2005). Machine learning techniques have started to produce surprising results for nonlinearity compensation in the field of optical communications (AMARI *et al.*, 2019).

It is important to mention that machine learning methods are beginning to be employed in fiber network analysis (MATA *et al.*, 2018), ranging from network planning in the physical layer using genetic algorithms (MORAIS *et al.*, 2011) to estimating quality of transmission (AZODOLMOLKY *et al.*, 2011), virtual topologies, intra-datacenter networking, and automating network management operations. In this work, we also apply machine learning methods in this context: our proposed approach involves using neural networks to recover signals degraded by spreading spots in the constellation.

In a QAM signal (see Appendix B), decreasing separation between adjacent symbols makes it more difficult for the receiver to properly decode the signal. Then, we are recovering the transmitted signal that reaches the receiver at the end of the optical link. Figure 1 a) shows the transmitted 16-QAM signal at the start of the 135-kilometer optical link, while Figure 1 b) shows the received signal, which is affected by noise and nonlinear distortion introduced by the long-haul network.

Figure 1 – 16-QAM a) transmitted and, b) received signal



Source: Author

1.2 Contribution

The main contribution of this work is the design of a nonlinear equalizer using neural networks, considering not only the nonlinear compensation capability but also the computational complexity. This is important because increasing the number of processed symbols naturally increases the complexity during the training stage.

Furthermore, to achieve this contribution, we addressed the following:

I) Better understanding of the interaction between chromatic dispersion and the Kerr effect. When considering information from multiple adjacent symbols, performance improves, but the complexity of the equalizer increases. In other words: the design becomes more challenging.

II) Improvement of software performance for code parallelization and reduced computation time. *Python*, the programming language used in this work, is powerful by default but typically uses only one computation core. We successfully parallelized the computations, which is not trivial since *Python* is not fundamentally designed for multiprocessing. Furthermore, we present a detailed analysis of the different results obtained for each configuration at the end of this work, and that includes the computational complexity results.

1.3 Work structure

This text is structured as follows: in Chapter 2, we introduce the main concepts of optical telecommunication signals and artificial neural networks that are essential for understanding the final results, explaining in detail the chromatic dispersion and the Kerr effect. Then, we introduce the current Artificial Intelligence (AI) scenario, followed by the artificial neuron and the activation functions. Furthermore, in Chapter 3, we present the approach used to solve the problem and compare the results for the Bit Error Ratio (BER) across different models. Finally, in Chapter 4, we present the conclusions of this work.

2 BACKGROUNDS AND METHODS

In this section, we present a literature review to better understand the problem approach and results. Although the reader may be familiar with optical communication systems, he may not be familiar with AI concepts and neural networks. This is why we provide a detailed description of both areas in this chapter.

2.1 Nonlinear intersymbol interference

In the following subsections, we are going to present concepts related to optical communication systems that are needed for a better understanding of this text. These concepts are the chromatic dispersion and the Kerr effect. We will also explain the interaction between these two optical system impairments.

2.1.1 Chromatic dispersion

An electromagnetic wave travels at constant speed in vacuum for all frequency components, but that does not happen when the wave travels in a dieletric medium. When that happens, the velocity of the wave is dependent on the wavelength. Then, different frequency components will arrive at different timesteps at the end of the medium, leading to the so-called “Chromatic Dispersion”. There are two main contributions for chromatic dispersion. The first contribution is the Material dispersion, and this contribution is intrinsic to the molecular structure of the material that was used to create the fiber. This dispersion is produced by the response of the bound electrons inside the dieletric medium, which oscillate with different amplitudes depending on the frequency of the incident electromagnetic wave. The refractive index can be approximated by the Sellmeier equation (KARMAKER, 2009):

$$n^2 = 1 + \sum_{j=1}^{m} \frac{B_j \omega_j^2}{\omega_j^2 - \omega^2}. \tag{2.1}$$

B_j is the strength of the jth resonance and w_j is the resonance frequency. The propagation constant β is dependent on wavelength:

$$\beta = \frac{\omega}{c} n(\omega), \tag{2.2}$$

where c is the speed of light in a vacuum. The second contribution is waveguide dispersion. The effective index of the mode depends on the fraction of power in the core and cladding at a particular wavelength. Even if the refractive indices of the core and cladding are assumed to be independent of wavelength, the effective index still changes with wavelength. Waveguide dispersion depends on fiber design parameters, such as core radius and core-cladding index difference (AGRAWAL, 2000).

2.1.2 Kerr effect

The Kerr effect is related to the dependence of the optical fiber's refractive index on the transmitted signal power. This effect is considered a limitation in Dense Wavelength-Division Multiplexing (DWDM) systems (see Appendix C) (AMARI *et al.*, 2017). In DWDM systems with multiple channels, compensation of the Kerr effect reduces nonlinear interactions between channels, which is very important to decrease noise. Fiber propagation affected by the Kerr effect is modeled using the nonlinear *Schrödinger* equation for single-channel systems and coupled nonlinear *Schrödinger* equations for DWDM systems. Among nonlinear effects, the Kerr effect has a significant impact on channel capacity (KAHN; HO, 2004). It leads to SPM, XPM and FWM. These effects limit the number of available channels (SOMAN, 2021). Let us explore each of these impairments in more detail:

I) Self-Phase Modulation: SPM refers to the self-induced phase shift caused by the optical field as it propagates through the optical fiber. Higher intensity parts of the optical pulse experience a higher refractive index in the fiber compared to lower intensity parts. This results in a temporal variation of the refractive index. And that leads to a temporal variation in the phase. The nonlinear phase shift is represented in Equation 2.3:

$$\phi_{NL} = \int_0^L P(z)dz = \gamma P_{in} L_{eff}, \tag{2.3}$$

where $P(z) = P_{in}e^{(-\alpha z)}$ accounts for fiber losses and L_{eff} is the effective interaction length, defined as:

$$L_{eff} = \frac{1 - e^{-\alpha L}}{\alpha}, \tag{2.4}$$

noting that, L is the fiber length, P_{in} is the input power and γ is the nonlinear coefficient. Also, α represents the loss coefficient.

II) Cross-Phase Modulation: XPM occurs because the nonlinear refractive index encountered by an optical beam also depends on the intensity of copropagating beams. This effect introduces distortion in the pulse and causes spectral broadening.

The phase shift for the jth channel is:

$$\phi_{NL}^{j} = \gamma L_{eff} \left(P_j + 2 \cdot \sum_{m \neq j} P_M \right), \tag{2.5}$$

where P_j is the optical power of the channel j experiencing SPM and P_M represents the optical power of the all other channels contributing to XPM.

III) Four-Wave Mixing: FWM arises due to the nonlinear response of bound electrons in a material to an optical field. The polarization induced in the medium contains both linear and nonlinear terms, determined by susceptibilities of different orders (SHRADDHA; VIKAS; SHANTANU, 2016).

The phase mismatch can be written as:

$$\Delta = \beta(\omega_3) + \beta(\omega_4) - \beta(\omega_1) - \beta(\omega_2), \tag{2.6}$$

where $\beta(\omega)$ is the propagation constant for an optical field with frequency ω.

One remark before we proceed: in a system with a single polarization and one carrier, the signal is only affected by SPM. In dual-polarization systems, the signal is influenced by both XPM and FWM. If we consider only one polarization in a dual-polarization system, we can mitigate only SPM. Nevertheless, by considering both polarizations, we can mitigate both SPM and XPM. This mitigation involves a trade-off between complexity and performance (COSTA, 2022). For an optical link with a single carrier and no repeaters, nonlinear effects arise from interactions between adjacent symbols. These effects will be explained in more detail in Section 2.1.3.

2.1.3 Interaction between chromatic dispersion and Kerr effect

In a dual-polarization system, neglecting the phase noise, the complex envelope of "X" polarization is given by:

$$A_x(t) = \sum_{k=-N}^{N} A_{x,i}^{(i)}\, f_p(t - k\,T_S) + j \sum_{k=-N}^{N} A_{x,q}^{(i)}\, f_p(t - k\,T_S). \tag{2.7}$$

And, for the "Y" polarization we have:

$$A_y(t) = \sum_{k=-N}^{N} A_{y,i}^{(i)}\, f_p(t - k\,T_S) + j \sum_{k=-N}^{N} A_{y,q}^{(i)}\, f_p(t - k\,T_S). \tag{2.8}$$

Where f_p represents the pulse shape, T_S the symbol period and N the number of symbols.
The vectorial analytic modulated signal can be written as:

$$\vec{A}(t) = \begin{pmatrix} \hat{x} & \hat{y} \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix} \exp(j\omega_0 t). \tag{2.9}$$

The fiber field is governed by the following equations:

$$\frac{\partial A_x}{\partial z} + \beta_{1x}\frac{\partial A_x}{\partial t} + \frac{j\beta_2}{2}\frac{\partial^2 A_x}{\partial t^2} + \frac{\alpha}{2}A_x = j\gamma\left(|A_x|^2 + \frac{2}{3}|A_y|^2\right)A_x + \frac{j\gamma}{3}A_x^* A_y^2 \exp(-2j\Delta\beta z), \tag{2.10}$$

and, for the other polarization:

$$\frac{\partial A_y}{\partial z}+\beta_{1y}\frac{\partial A_y}{\partial t}+\frac{j\beta_2}{2}\frac{\partial^2 A_y}{\partial t^2}+\frac{\alpha}{2}A_y=j\gamma\left(|A_y|^2+\frac{2}{3}|A_x|^2\right)A_y+\frac{j\gamma}{3}A_y^*A_x^2\exp(+2j\Delta\beta z). \quad (2.11)$$

Where β_2 is the second order dispersion term that leads to pulse spreading and β_1 represents the group velocity term. In the regime of weak nonlinearity, perturbation theory can be used to approximate the nonlinear phase shift, ϕ_{NL}, at the output of an optical fiber of length L. This phase shift can be expressed in terms of the optical amplitude $A(z,t)$, resulting in the following form:

$$\phi_{NL}(L,t)=\int_0^L \gamma|A(z,t)|^2 dz. \quad (2.12)$$

And, $\gamma = 1.3\ \mathrm{W}^{-1}\mathrm{km}^{-1}$. Assuming the fiber's nonlinear effects are sufficiently weak, we can consider chromatic dispersion as the primary mechanism influencing the evolution of the optical amplitude. Then, the optical amplitude can be approximated by (AGRAWAL, 2000):

$$A(z,t)\approx A(0,t)*h_f(z,t). \quad (2.13)$$

$h_f(z,t)$ is the linear impulse response of the fiber (SAVORY, 2008), and $A(0,t)$ is the input amplitude composed of $2N+1$ symbols, which we can write as:

$$A(0,t)=\sum_{k=-N}^{N} A_k f_p(0,t-k\,T_S). \quad (2.14)$$

A_k denotes the complex amplitude associated with a specific constellation point, while $f_p(0,t-kT_\mathrm{S})$ represents the pulse shape function of the kth symbol, centered at $t=kT_\mathrm{S}$. Given the symmetric spreading of the pulse, the optical amplitude at the output, evaluated at the center of the symbol of interest, is:

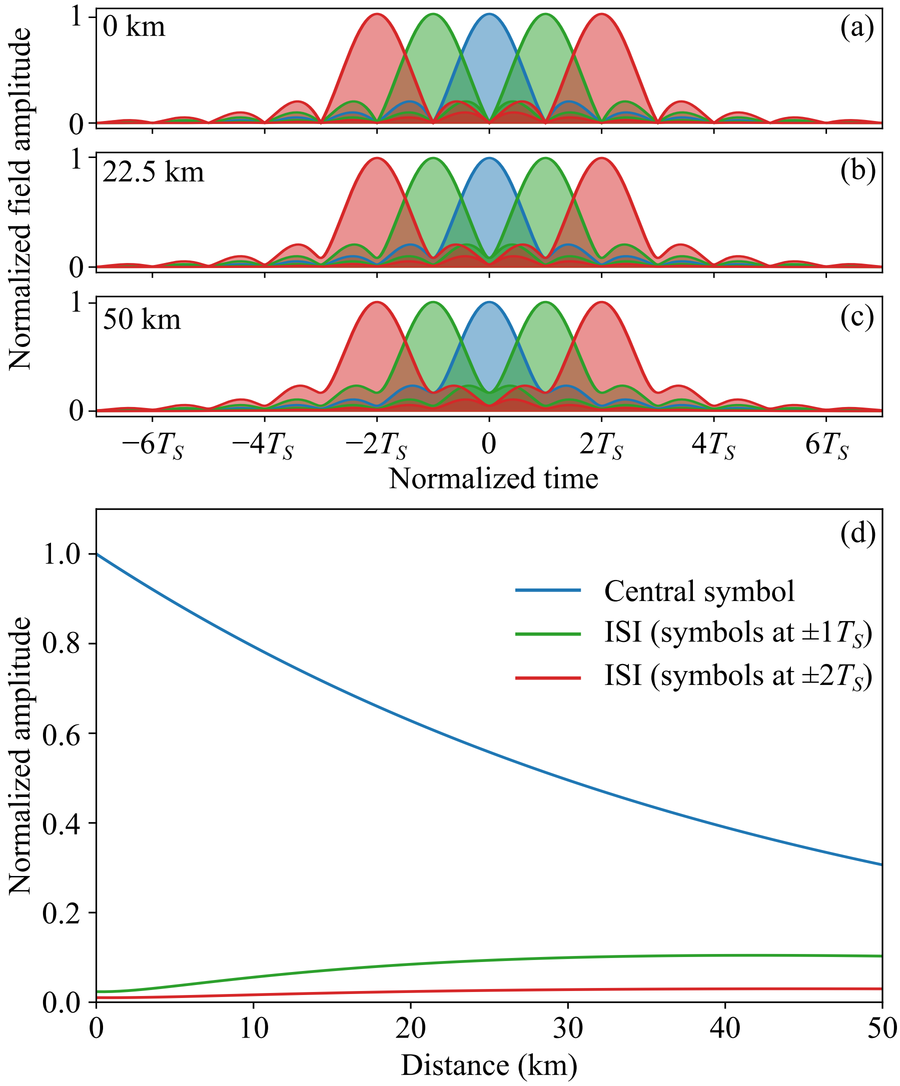
$$A(L,0)\approx A_0\left[f_p(0,t)*h_f(L,t)\right]_{t=0}+\sum_{\substack{k=-N\\k\neq 0}}^{N} A_k\left[f_p(0,t-kT_S)*h_f(L,t)\right]_{t=0}. \quad (2.15)$$

In the case where the amplitude of the ISI is much smaller than that of the symbol of interest, the optical intensity can be approximated by:

$$\begin{aligned}|A(L,0)|^2 \;\approx\;& |A_0|^2\cdot\left|\left[f_p(0,t)*h_f(L,t)\right]_{t=0}\right|^2 \quad (2.16)\\ &+2\sum_{\substack{k=-N\\k\neq 0}}^{N}\left(\left|A_0\left[f_p(0,t)*h_f(L,t)\right]_{t=0}\right|\cdot\left|A_i\left[f_p(0,t-k\,T_S)*h_f(L,t)\right]_{t=0}\right|\cos(\phi_j)\right).\end{aligned}$$

The first term on the right-hand side represents the contribution of the symbol of interest. The second term represents the intensity contribution, produced by the interference between the symbol of

Figure 2 – Pulse spreading considering rise cosine pulses at 14 GBaud for a) 0km, b) 22.5km, c) 50km and d) amplitude of the pulse at optimum sampling instant



Source: Author

interest and its neighboring symbols. This contribution is influenced by the phase difference between the interfering terms, ϕ_j, and exhibits a complex relationship with the complex amplitudes of the symbols and the dispersion properties of the fiber.

Figure 2 (a), (b), and (c) illustrates the shape of five symbols at the fiber input. The analysis assumes 14 GBaud pulses with a roll-off factor of 0.2 and an SSMF with parameters specified in a previous work (COSTA *et al.*, 2023). These subfigures qualitatively demonstrate the overlapping of pulses at the center of the symbol of interest. To quantify this overlap, Figure 2 (d) presents the evolution of the amplitude of the central pulse at its peak, together with the contributions from adjacent pulses during propagation. It can be seen that the amplitude of the ISI terms starts low but increases until it becomes comparable to the contribution of the central pulse. Here, it is important to recognize one thing: the final impact on SPM will depend on the relative phases. These results suggest that adjacent symbols exhibit significant overlap at the center of the symbol, contributing to nonlinear phase modulation.

To analyze the contribution from adjacent symbols, we used VPI Transmission Maker. Our data corresponds to 16-QAM signals with input power ranging from 0 dBm to 12 dBm for various fiber lengths. And we have the in-phase and quadrature values for both the transmitted and received signals. Part of our approach involves recovering the signal distorted by the harsh, noisy telecommunication channel and mitigating the nonlinear distortion. We try to achieve a lower BER using the neural network method in contrast to the maximum likelihood method. The dataset corresponds to a signal transmitted through an optical coherent digital system, using standard single-mode fiber with a length of 135 kilometers, without repeaters and with dual-polarization. A remark of extreme importance: it is not the goal of this work to describe the simulation setup in detail, as this has already been extensively covered in previous works (COSTA, 2022).

In this work, we propose using neural networks to compensate the nonlinear effects. To better understand this approach, it is essential to first introduce some concepts of AI relevant to this process.

2.2 Compensation using neural networks

We cannot explain how the brain performs its various mental skills. Nevertheless, scientists and engineers are able to produce machines that can perform certain types of work that would require mental skills, such as: voice recognition (ORUH; VIRIRI; ADEGUN, 2022) and writing recognition (MEIßL *et al.*, 2022). We can trace in the literature surprising results even in music sheet recognition (DUA *et al.*, 2020). AI is the name given to that research (MINSKY, 1990). Furthermore, there are multiple definitions that can be used to define AI. In the words of Charniak and McDermott: “The study of mental faculties through the use of computational models”, and “The art of creating machines that perform functions that require intelligence when performed by people”. It is worth mentioning that AI studies started after World War II. AI at the present moment that this work is elaborated can enable computers to write poetry, write texts and even play chess (RUSSELL; NORVIG, 2016). Alan Turing, (who is considered to be the founding father of computer science and AI) proposed what is called the “*Turing test*” back in 1950. The “imitation game” proposed by Turing consists of testing the computer’s ability to mimic a human, judged by another human being who could not see the machine but could ask questions. The computer is able to pass the test if the human is not able to tell if the answers come from a computer or from another human.

The first years of AI study were successful, considering that there were limited programming and computing tools at the time (from 1952 to 1969). In 1952, Arthur Samuel wrote a computer program that was able to play checkers (SCHAEFFER; LAKE, 1996), and the program was demonstrated on television in 1956, producing an impressive reaction. This reaction was later amplified, when in 1997, the chess computer named *Deep Blue* defeated Russian chess champion Garry Kasparov. The experience was later described by Kasparov in his book, “*Deep thinking: where machine intelligence ends and human creativity begins*” (KASPAROV, 2017): “When I sat down at the chess board across from Deep Blue, I immediately sensed something new, something unsettling...” (KASPAROV, 2017, p.13), and: “I couldn’t help wondering, Was it invincible? Was my beloved game of chess over?” (KASPAROV, 2017, p.124).

In more recent years, the passing of the “Turing test”, has become a parameter of concern, as technology leaders and researchers of AI recognize that these systems may present a potential threat to humanity.

In 2023, Yoshua Bengio participated in an open letter asking to slow down the development of AI systems that can currently pass the Turing test, and thus "trick a human being into believing it is conversing with a peer rather than a machine". Along with Yoshua, technology leaders, including Elon Musk and Apple co-founder Steve Wozniak have called for a pause in the training of any "AI system more powerful than GPT-4" (EURONEWS, 2023). It states that AI could potentially change the trajectory of life on Earth, and it should be planned carefully. It is not surprising that Alan Turing already expressed concern about this topic long before this open letter came into existence. In his own words:

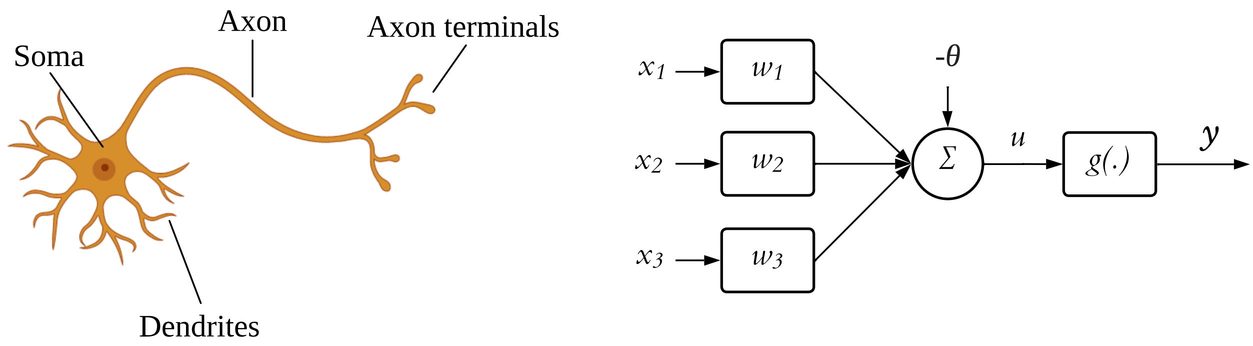
> There would be plenty to do in trying, say, to keep one's intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control (TURING, 1996, p.4).

2.2.1 Artificial neuron

The brain structure was changed during evolution, but the properties of a neuron remain the same across the animal kingdom: the capability to accept signals, excitability and transportation of electrical messages to chemical output (SANDLER; TSITOLOVSKY, 2008). The Artificial Neural Network (ANN) is a type of supervised learning approach in which there is an emulation of the biological neural network. The first article related to this emulation, called "A logical calculus of the ideas immanent in nervous activity" dates back to 1943, where the authors presented the first mathematical model based on the biological neuron, *i.e.*, the first-ever produced artificial neuron model (MCCULLOCH; PITTS, 1943).

The human brain is composed of a large number of neurons, linked to each other by a large amount of synapses (LIN, 2017). A great effort in neuroscience is focused on understanding the brain's learning process. Many mechanisms have been proposed to explain synapse activity (SCHMIDGALL *et al.*, 2023). Synapse formation involves two agents: axons and dendrites. The axon of a presynaptic neuron needs to be guided to make synapses with the correct target, *i.e.*, the dendrites (usually), of postsynaptic neurons (JAN; JAN, 2001). ANNs are built from analogy of the biological neural networks. ANNs follow closely the model of the brain, and that is why the terminology has everything to do with the neuroscience terminology (DONGARE *et al.*, 2012). We now know, through neuroscience that dendrites are the receivers and the axons the transmitters (MARGRIE; URBAN, 2007). Figure 3 illustrates the comparison, between the structure of a biological neuron and an artificial neuron.

Figure 3 – Biological and artificial neuron



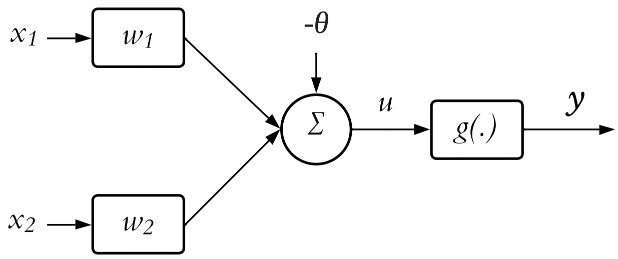
Source: SILVA, 2010

The neuron receives input from other neurons but information will only pass to other neurons if a certain threshold is achieved. In the case of an artificial neuron this threshold value depends on activation function, bias, weights and input values. On the biological neuron only if certain voltage difference value is achieved the neuron depolarization occurs, after which neurotransmitter molecules bind to receptor on neuron dendrites and the ion channels are opened (LIBRETEXTS, 2021).

Now, our interest is to model the neural network as a mathematical model of the “Biological” neuron. The effect of the synapse is represented by the connection weight that modulates the effect of the input signal, and the nonlinear aspect is represented by a transfer function. Learning capability is dependent on the weight adjustment (GUBRAN, 2013).

Let us have a closer look at the mathematical analysis of the Perceptron. If we have one Perceptron with two inputs, as shown in Figure 4 the output can be described by Equation 2.17 (SILVA; SPATTI; FLAUZINO, 2010).

Figure 4 – Neural network training process



Source: SILVA, 2010

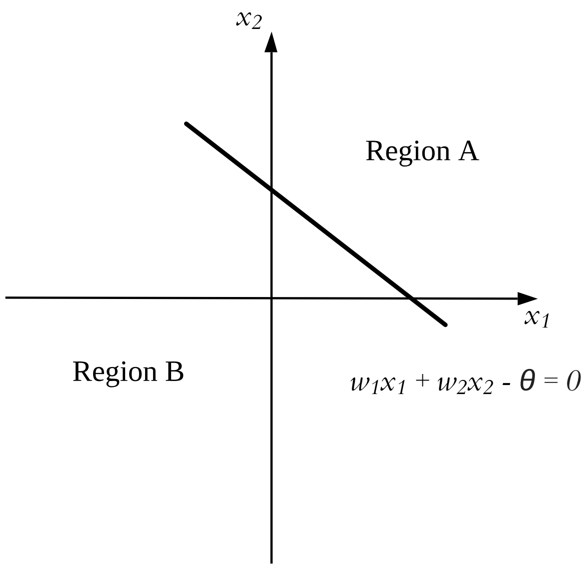
$$y = \begin{cases} 1, & \text{if } \sum w_i x_i - \theta \geq 0 \quad \Rightarrow \quad w_1 x_1 + w_2 x_2 - \theta \geq 0, \\ -1, & \text{if } \sum w_i x_i - \theta < 0 \quad \Rightarrow \quad w_1 x_1 + w_2 x_2 - \theta < 0. \end{cases} \tag{2.17}$$

The inequalities from Equation 2.17 are represented by a linear expression, in which y represents the neural network output, x_i represents the neural network inputs, w_i represents the weights, θ the bias term and $g(.)$ the activation function, which we are going to describe in more detail in Section 2.2.2. The decision boundary is then defined by:

$$w_1x_1 + w_2x_2 - \theta = 0. \tag{2.18}$$

Figure 5 illustrates the decision boundary. The Perceptron can only separate linearly separable classes.

Figure 5 – Perceptron decision boundary

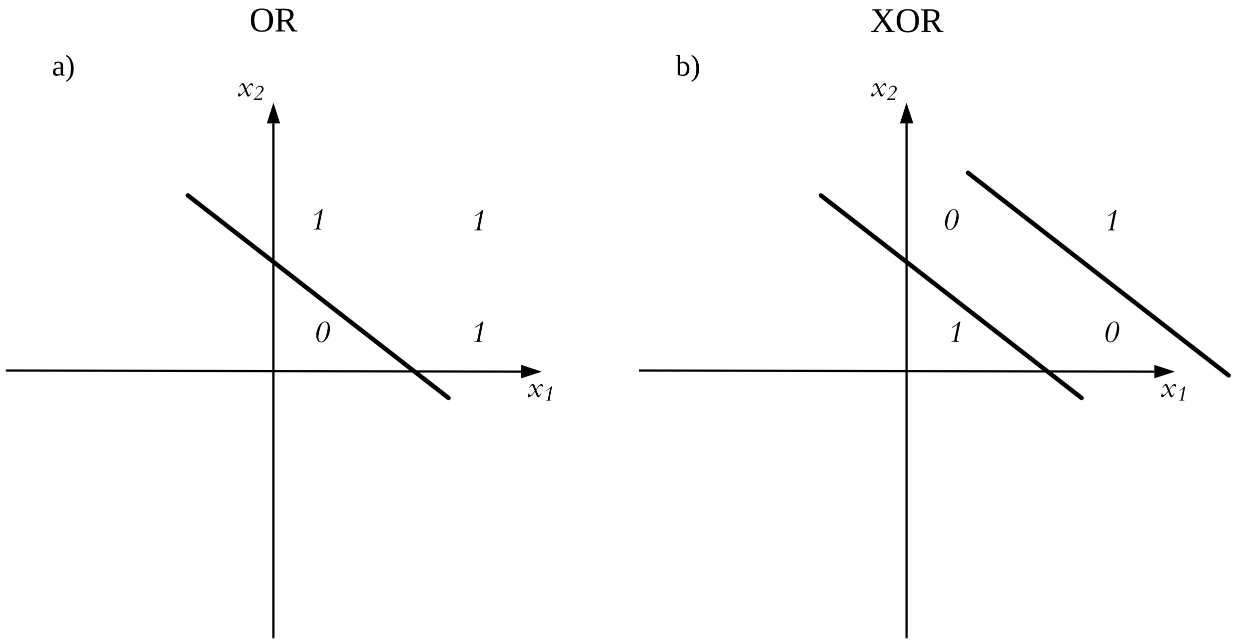


Source: SILVA, 2010

A Perceptron composed of one layer can be used as an classifier only in problems in which the problem classes can be mapped as linearly separables. This principle comes from the "Perceptron Convergence Theorem" (MINSKY; PAPERT, 1969).

Due to Minsky and Papert book "Perceptron", the interest in neural networks back in 1969 was greatly diminished, as it was proved that a model like Rosenblatt's network model, (ROSENBLATT, 1958) *i.e.*, the one-layer Perceptron was incapable of learning certain simple boolean functions. We know now that, even though it may not always be feasible, neural networks can approximate any smooth function to any desired accuracy using only one hidden layer (LIN; TEGMARK; ROLNICK, 2017). For that reason, it is important to introduce more layers and activation functions when describing the model of a neural network. Multilayer neural networks became public by the end of the 1980s, mainly after the publication of the book *Parallel Distributed Processing* (RUMELHART *et al.*, 1986). Figure 6 illustrates the linearly separable problem, including the boolean operation "XOR" which is not linearly separable, that is: not possible to be learned by a single layer Perceptron.

Figure 6 – The "XOR problem"



Source: Author

We can describe mathematically the "MultiLayer Perceptron" (MLP) by the following equations:

$$I_j^{(1)} = \sum_{i=0}^{n} W_{ji}^{(1)} x_i \quad \Rightarrow \quad I_j^{(1)} = W_{j0}^{(1)} x_0 + W_{j1}^{(1)} x_1 + \ldots + W_{jn}^{(1)} x_n \tag{2.19}$$

$$I_j^{(2)} = \sum_{i=0}^{n_1} W_{ji}^{(2)} Y_i^{(1)} \quad \Rightarrow \quad I_j^{(2)} = W_{j0}^{(2)} Y_0^{(1)} + W_{j1}^{(2)} Y_1^{(1)} + \ldots + W_{jn_1}^{(2)} Y_{n_1}^{(1)} \tag{2.20}$$

$$I_j^{(3)} = \sum_{i=0}^{n_2} W_{ji}^{(3)} Y_i^{(2)} \quad \Rightarrow \quad I_j^{(3)} = W_{j0}^{(3)} Y_0^{(2)} + W_{j1}^{(3)} Y_1^{(2)} + \ldots + W_{jn_2}^{(3)} Y_{n_2}^{(2)}, \tag{2.21}$$

where $W_{ji}^{(l)}$ are matrices of weights which elements corresponds to the synaptic weight connecting j neurons from layer l to the ith neurons of layer $l-1$. $I_j^{(l)}$ are vectors which elements corresponds to the input taking into account the jth neuron from layer l.

Finally, $Y_j^{(l)}$ are vectors which elements corresponds to the output of the jth neuron that is connected to layer l, defined as:

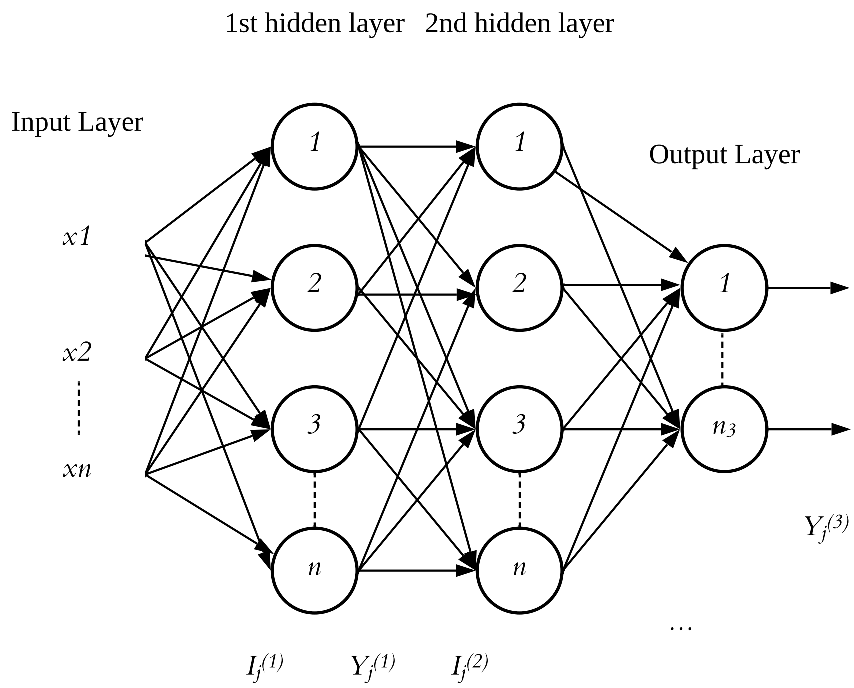
$$Y_j^{(1)} = g(I_j^{(1)}), \tag{2.22}$$

$$Y_j^{(2)} = g(I_j^{(2)}), \tag{2.23}$$

$$Y_j^{(3)} = g(I_j^{(3)}). \tag{2.24}$$

Figure 7 illustrates a MLP.

Figure 7 – The multilayer perceptron topology considering two hidden layers



Source: SILVA, 2010

Let us take a closer look now at what is called the "training" process. One vital concept behind the machine learning algorithms is showing a great number of examples to the computer. In 1949 the first training method was informed by the canadian psychologist Donald Olding Hebb. After his name came the term "Hebbian learning" (HEBB, 2005). Using his words: "The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become associated so that activity in one facilitates activity in the other."(HEBB, 2005, p.70).

We will now discuss the mathematical modelling of this learning process. Considering the kth training sample for the topology illustrated in Figure 7, the quadratic error function used to evaluate model performance is given by:

$$E(k) = \frac{1}{2}\sum_{j=1}^{n_3}(d_j(k) - Y_j^{(3)}(k))^2, \tag{2.25}$$

where $Y_j^{(3)}$ is the result produced by the jth neuron on the output of the network taking into account the kth training sample and $d_j(k)$ is the respective expected value.

The training process objective is to adjust the weight matrices to $W_{ji}^{(3)}$ to minimize the error between the output and the target output value. Considering the error value from Equation 2.25 in relationship with the kth sample from the jth neuron output and using the gradient definition with the chain rule we have:

$$\nabla E^{(3)} = \frac{\partial E}{\partial W_{ji}^{(3)}} = \frac{\partial E}{\partial Y_j^{(3)}} . \frac{\partial Y_j^{(3)}}{\partial I_j^{(3)}} . \frac{\partial I_j^{(3)}}{\partial W_{ji}^{(3)}} . \tag{2.26}$$

From equations 2.21, 2.24 and 2.25 we have:

$$\frac{\partial I_j^{(3)}}{\partial W_j i^{(3)}} = Y_i^{(2)}, \tag{2.27}$$

$$\frac{\partial Y_j^{(3)}}{\partial I_j^{(3)}} = g'(I_j^{(3)}), \tag{2.28}$$

$$\frac{\partial E}{\partial Y_j^{(3)}} = -(d_j - Y_j^{(3)}). \tag{2.29}$$

where g'(.) corresponds to the first order derivative of the activation function used. Combining:

$$\frac{\partial E}{\partial W_{ji}^{(3)}} = -(d_j - Y_j^{(3)}) \cdot g'(I_j^{(3)}) \cdot Y_i^{(2)}. \tag{2.30}$$

Then, weight matrix $W_{ji}^{(3)}$ should be made to minimize the error:

$$\delta W_{ji}^{(3)} = -\eta \cdot \frac{\partial E}{\partial W_{ji}^{(3)}} \Rightarrow \delta W_{ji}^{(3)} = \eta \cdot \delta^{(3)} \cdot Y_i^{(2)}, \tag{2.31}$$

where η is the learning rate and $\delta^{(3)}$ is the local gradient:

$$\delta^{(3)} = (d_j - Y_j^{(3)}) \cdot g'(I_j^{(3)}). \tag{2.32}$$

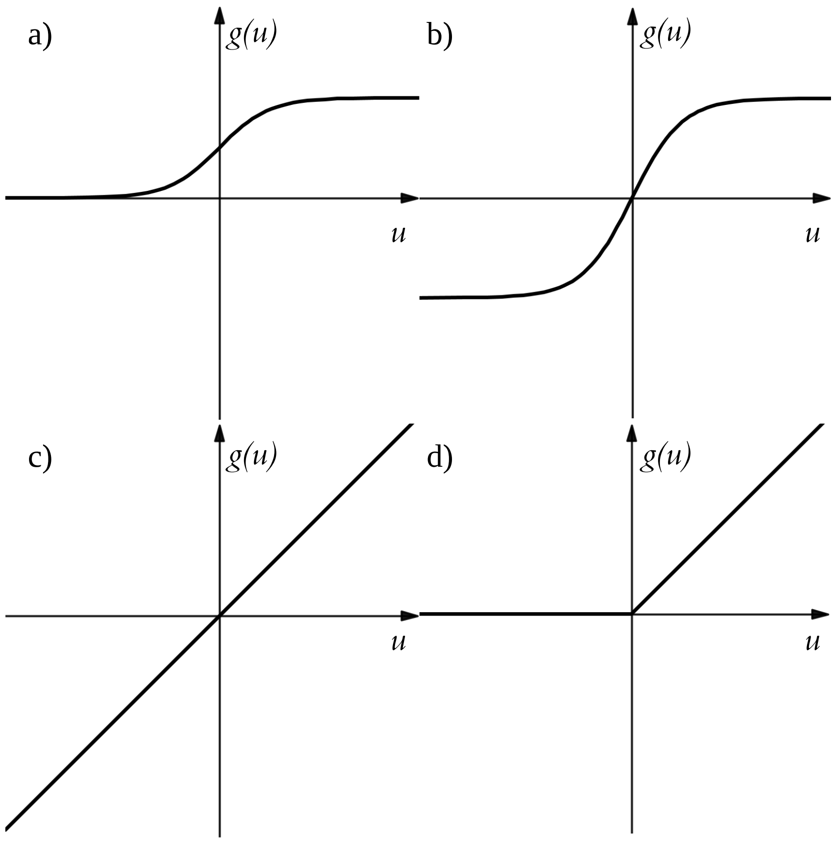
The neuron weights of the output layer are adjusted taking into account the difference observed between the produced output in comparison with the desired output values.

We can choose different activation functions according to the problem we are solving. As we mentioned, for regression problems we use linear outputs. For binary classifications sigmoid outputs can produce good results (ÇETIN; TEMURTAŞ; GÜLGÖNÜL, 2015) and for multiclass classification we can achieve good performance using softmax output activation functions (XIAO *et al.*, 2020). Activation functions will be described in the next section.

2.2.2 *Activation functions*

Now, we are going to talk about the activation functions. Activation functions can be divided as partially differentiable and totally differentiable functions. Partially differentiable functions are those which first order derivative points are non existant. Totally differentiable functions are those which first order derivatives exist and are known in all parts of the function domain. Let us focus on totally differentiable functions. The four main activation functions that belong to this group are: the logistic function, the hyperbolic tangent (tanh), the linear activation function and the rectified linear unit (ReLU).

Figure 8 – a) logistic, b) hyperbolic tangent, c) linear and d) ReLU math functions used as activation functions



Source: Author

a) Logistic function

The result produced from the output by applying the logistic function will always be between 0 and 1, having its mathematical expression written as:

$$g(u) = \frac{1}{1+e^{-bu}}. \tag{2.33}$$

The letter b is associated with the slope of the logistic function. Figure 8 - a) illustrates this function. Logistic function $g(u)$ values range from 0 to 1.

b) Hyperbolic tangent function

Let us introduce now the hyperbolic tangent function. The result of the output will always be between -1 and 1. The mathematical expression is:

$$g(u) = \frac{1 - e^{-au}}{1 + e^{-bu}}, \tag{2.34}$$

where a and b are scaling parameters that adjust the rate of change in the numerator and denominator, respectively. Figure 8 - b) shows the hyperbolic tangent function. Hyperbolic tangent function $g(u)$ values range from -1 to 1.

c) Linear function

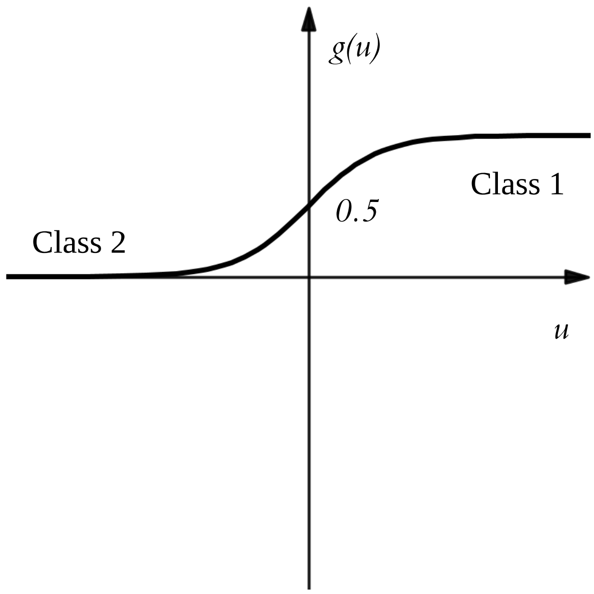
Let us introduce the linear function. The output values will be the same as the input. Figure 8 - c) shows a linear activation function (as we mentioned before, this activation function is used as the output for regression problems).

d) Rectified linear unit function

Finally, let us introduce the rectified linear unit. The ReLU activation function returns its argument u for positive values, and returns 0 for negative values. The derivative is equal to 1 when u is positive and 0 when negative. There are variations of the ReLU that can speed up convolutional neural networks convergence (JAAFARI; ELLAHYANI; CHARFI, 2021). Figure 8 - d) shows the ReLU function. ReLU is considered easy to calculate, with fast convergence speed and effectivity in reducing the vanishing gradient problem (LIN; SHEN, 2018)

The MLP can be used for regression problems and for classification. In the case of a binary classification problem, as mentioned in Section 2.2.1, sigmoid (also known as logistic) activation functions produce good results when used in the output layer. This can be more clear if we observe Figure 9 (AHMED, 2024). By defining the threshold at 0.5, we can achieve two very well defined output probabilities for a classification problem. Values above the defined threshold are going to be classified as *Class 1* and values below that threshold, classified as *Class 2*. In this work, we are dealing with a regression problem. That will be more clear in Section 2.3. What is important for now is to understand that it would not be appropriate for our regression problem to use activation functions well suited for classification problems, but these could be extensively used in different works that use the MLP. The activation function used in this work was the ReLU function, because of its fast convergence and ability to introduce nonlinearities, two characteristic that are very well suited for the aspect of this work.

Figure 9 – Logistic function for binary classification



Source: AHMED, 2024

2.3 Problem approach

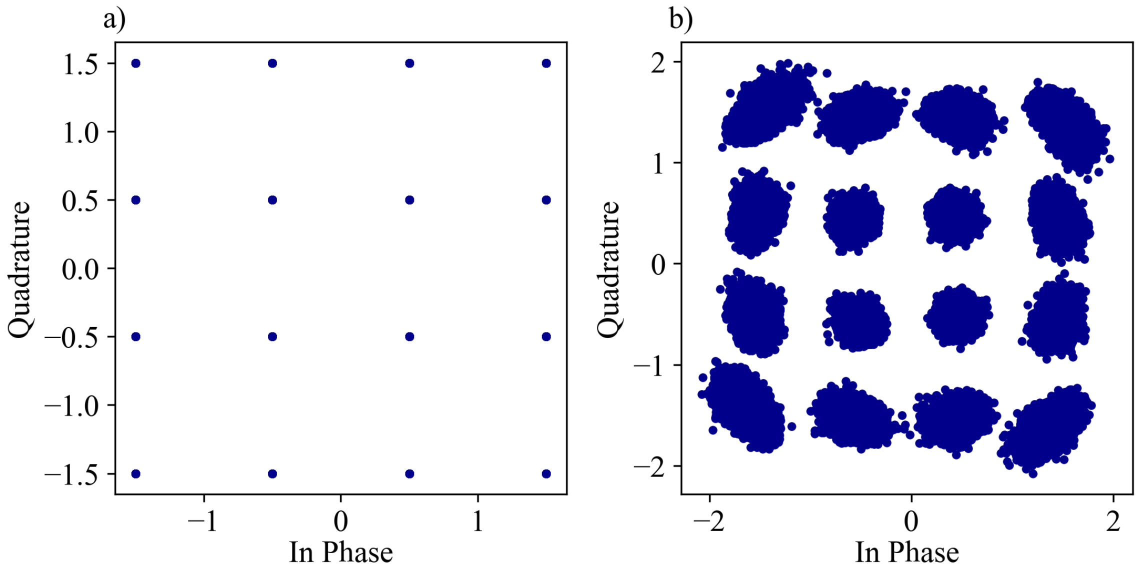
The neural network in this work was implemented using Python with *scikit-learn*, an open-source machine learning library. Released in 2007, it has become one of the most popular machine learning libraries for Python (PEDREGOSA *et al.*, 2011). Built as an extension to the SciPy library, scikit-learn is based on NumPy and Matplotlib and is popular in academic research due to its well-documented, easy-to-use, and versatile API (HACKELING, 2017).

One of the first steps in a machine learning modeling problem is to divide the dataset into training and testing samples. In statistics, it is known that the more data points we use, the more accurate the results we obtain (CHO *et al.*, 2015). However, in practice, we cannot always determine if the current model is adequate for the problem at hand. If we use all available data for training, we risk overfitting[1], which would also be a methodological mistake. In this work, we initially allocated 50% of the data for training and 50% for testing. Now, this was an arbitrary choice and we don't think it should have a significant impact on the results or research goals. Different studies in the field use varying training and testing ratios: commonly, 70% for training and 30% for testing, or 60% and 40%. Empirical analysis suggest that allocating about 80% of the data for training and 20% for testing is optimal, and there are detailed mathematical demonstrations supporting this ratio (GHOLAMY; KREINOVICH; KOSHELEVA, 2018), although there is no exact and definitive consensus in the machine learning area.

Figure 10 illustrates the transmitted (a) and received (b) optical signals for an input of 9 dBm.

[1] Overfitting occurs when a model learns patterns specific to the training dataset and fails to generalize to new data.

Figure 10 – 16-QAM a) transmitted and b) received constellation with 9 dBm input signal

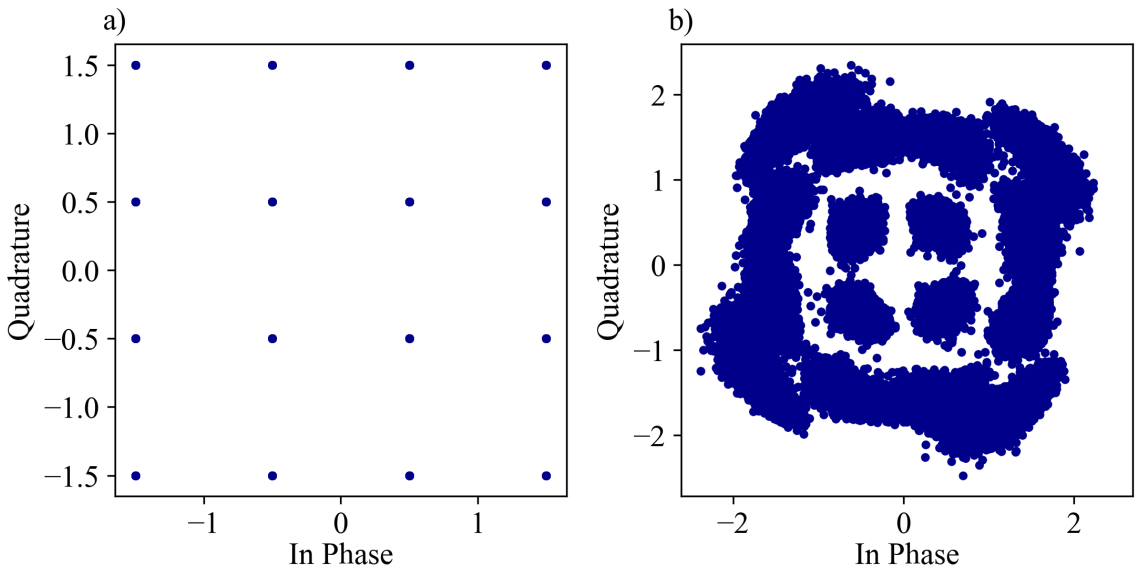


Source: Author

As illustrated in Figure 10, our received signal is affected by nonlinear distortion. And it is very different from the signal illustrated in Figure 1 in our introduction (the resulting distortion for an input signal power of 0 dBm).

Figure 11 illustrates the constellation received for 12 dBm. In this case, the rotation becomes even more noticeable.

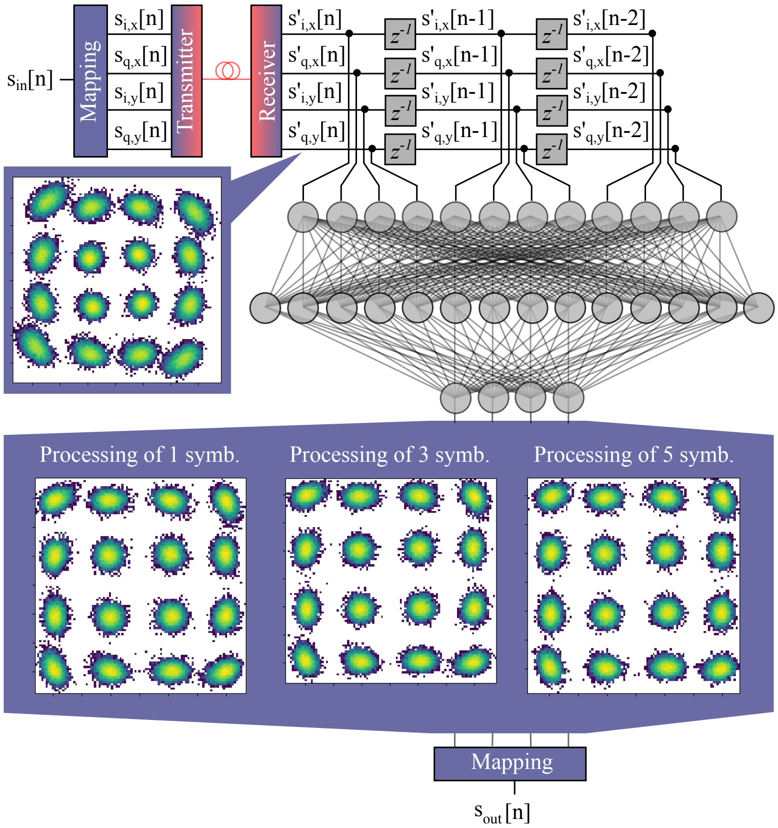
Figure 11 – 16-QAM a) transmitted and b) received constellation with 12 dBm input signal



Source: Author

That is the expected behavior: for increased power input we would see more pronounced points displacement in the output signal. The Kerr effect is a type of optical nonlinearity where the refractive

Figure 12 – Block diagram of the scheme for compensation of the nonlinear ISI



Source: Author

index of a material depends on the intensity of the light passing through it. As the input power increases, we see more changes in the phase velocity of light, leading to phase modulation. It is important to mention at this point: the performance for coherent systems takes into account the effect of noise and nonlinearities. The noise is stochastic, and we cannot compensate properly. Our task is restricted to the mitigation of the nonlinear effects introduced by the optical link (COSTA, 2022). Our goal is to mitigate that distortion, recovering the transmitted signal at the receptor with lowest BER, considering the complexity needed to perform this task (computationally).

Most machine learning techniques for compensation operate on received distorted symbols influenced by both the symbol itself and its neighboring ones. Now, to capture the effect of multiple symbols, a model with memory, such as a Recurrent Neural Network (RNN), can be utilized. But a memory-less approach like a MLP can also be used (where the model processes not just the target symbol but also several adjacent symbols). And that is exactly the method that is used in this work. We chose the latter approach for two main reasons:

a) the MLP is simpler to train and optimize, and,

b) it allows us to control the number of symbols interacting nonlinearly.

Figure 12 presents the block diagram of a phase and polarization diversity digital coherent system, illustrating how the MLP processes the in-phase and quadrature components of several adjacent symbols. To better demonstrate, the diagram includes configurations with 1, 3, and 5 adjacent symbols for simultaneous processing.

3 RESULTS AND DISCUSSION

This chapter will present the dataset used, how the neural network was trained and the results for different number of neurons, different number of processed symbols, complexity and number of operations. We are going to show the results comparing the BER and the model computational cost.

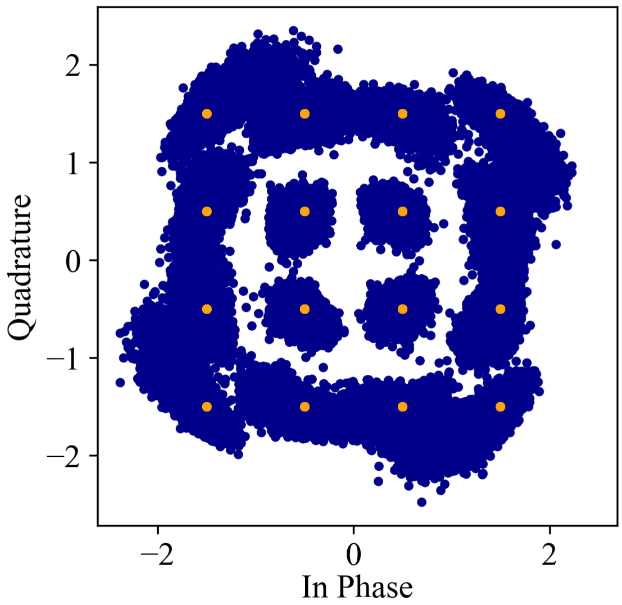
3.1 Dataset

The dataset used in this work is composed of 16-QAM signals. Signals modulated using this modulation scheme have 2 dimensions, and we also have the dual-polarization: the in-phase and quadrature values for the "X" and "Y" polarizations. In this work, we propose to mitigate the nonlinearity introduced by the channel using neural networks. Then, when modelling the problem it is important to understand which are the inputs and outputs of the neural network.

For machine learning problems, our intent is to provide to the network data for inputs and outputs, as pointed out in Section 2.2.1. We present a great number of examples to the computer, and through the training process the model will be able to approximate values for the output at a given input, based on previous values that were presented to the model during the training process. Here, we are introducing to the neural network values of both polarization in-phase and quadrature values, before and after the passing of the signals through the optical link. After the NN was trained with a great number of examples, it may be able to approximate the values for the signal amplitude in the receiver output from the optical link provided inputs.

This can be more clear if we overlap the constellations presented in Figure 10. Our objective here is to mitigate the spreading spots and rotation of the constellation that represents our 16-QAM signal. The signal at the beginning of the optical link is represented by the points -1.5, -0.5, 0.5 and 1.5, represented by orange in Figure 13, while, the points at the end of the optical link can assume any value around this points, possibly even higher than 2 or lower than -2 depending, of course, on how strong is the nonlinear distortion introduced by the channel.

Figure 13 – Inputs and outputs for neural network modelling



Source: Author

Now, in machine learning problems the input is represented by the data provided to the model. In this work, we are modelling a nonlinear equalizer. Then, it should be clear that the input for an equalizer is going to be the distorted signal. The output should be the equalized signal, with values as close as possible to the points represented in orange. It should, at this point also be clear that, roughly, the closest the model output values are to the orange points with the correct in-phase and quadrature association, the lowest the BER will be.

One more way of visualizing our problem approach is by comparing the values of the input and the "*ideal output*" of our neural network. We say ideal output, because we expect the values to be as close as possible to this values. But in reality they will never exactly be. Let us take for instance the values for a Launched Optical Power (LOP) of 9 dBm in Table 1. These are the values that we are going to transmit through the fiber. Then, let us compare in Table 2 the values detected by the receptor at the end of the channel link.

Table 1 – Values for 9 dBm signal at the transmitter

In-phase X	In-phase Y	Quadrature X	Quadrature Y
1.5	-1.5	1.5	1.5
-0.5	1.5	-0.5	-0.5
1.5	-1.5	-0.5	1.5
1.5	-1.5	1.5	-0.5

Source: Author

We are going to present the data from both these tables to the neural network for training. Then, let us take our own remark at Section 2.2.1: "One vital concept behind the machine learning algorithms is showing a great number of examples to the computer". We are presenting these values to the network

Table 2 – Values for 9 dBm signal at the receptor

In-phase X	In-phase Y	Quadrature X	Quadrature Y
1.3664	-1.3855	1.4207	1.4791
-0.7212	1.3733	-0.5607	-0.6120
1.1047	-1.7917	-0.4180	1.445
1.3690	-1.5836	1.3914	0.7076

Source: Author

that is, through the learning process, i.e., training, capturing the patterns that govern the relationship between the ideal values (values at the input of the optical link: Table 1) and the values distorted by the telecommunication channel (the output values at the end of the optical link: Table 2). Then, it is expected that when presented only with input values, that is, distorted values at the end of the optical link, our trained machine learning model is able to estimate accurately the values that are expected to be at the input of our optical link, that is, the nonlinearity mitigated 16-QAM signals with values as close as possible to -1.5, 1.5, -0.5 and 0.5. And to avoid confusion, let us make one more important remark at this point: the input of our neural network is the output of our optical link system (i.e., the distorted 16-QAM signal), and the output of our neural network is our target value: the equalized 16-QAM signal at the input of our optical link.

Here, it should be clear the relationship between the ideal output and the input of the neural network. Equation 2.25 is used to evaluate the model performance by calculating the quadratic error between the ideal value and the value acquired as result of the neural network output. Let us assume one of the inputs of our neural network is the first row from Table 2. The neural network receives as input four values. We are also presenting the network the ideal output values, that is, the first row from Table 1. Now, let us assume the neural network during the training stage outputs the values exibihited in Table 3.

Table 3 – Possible values calculated by the neural network

In-phase X	In-phase Y	Quadrature X	Quadrature Y
1.5122	-1.5432	1.5923	1.4923

Source: Author

Equation 2.25 then becomes:

$$E = \frac{(1.5 - 1.5122)^2 + [-1.5 - (-1.5432)]^2 + (1.5 - 1.5923)^2 + (1.5 - 1.4923)^2}{2} \tag{3.1}$$

$$E = 0.005296. \tag{3.2}$$

From here, it should be more clear what the neural network is doing during the training process. It is using the data present in the training dataset to minimize this error calculated in 3.1 and 3.2 by adjusting the synaptic weights illustrated in Figure 7.

Here, again, we recall one goal of this work: while there are studies that do use machine learning models for nonlinearity mitigation, they only take into account the symbol of interest for that process. That would be equivalent to using as input of the neural network one single row of Table 1 and Table 2, represented by the one-dimensional matrix, i.e., vector: 1×4. This is not the case here. In this work we are going to use multiple symbols for the training process. More than that, we are going to vary the number of symbols to verify which combination is more computationally effective and/or results in best performance. This will be explored in the next section. But before that one last remark: we are going to use as input of our neural network combined rows of Table 1 and Table 2. Furthermore, if we use a number of processed symbols of 3, we are going to use for input of our NN blocks of 3 rows from our tables, represented by the matrix 3×4. In this work, we compared the results for number of processed symbols ranging from 1 to 15.

Many methods for nonlinearity compensation using machine learning contemplate only the symbol of interest. In this work, we propose to take into account adjacent symbols. This approach has the intent to “trick” the neural network, presenting to it during the training process different symbols, resulting in a form of recurrence for the network. This should not be confused with a RNN. RNNs, do have memory, but here we are using a MLP. The difference from the majority of other works that use it for nonlinear mitigation, is that in this work we are using multiple symbols for the development of the nonlinear equalizer. This, will increase the complexity of the training process, and for that reason one important goal of this work, is to also calculate the computational complexity of this equalizer (by the way, also reducing the computation time by parallelizing the code). Results comparing the BER and the computational cost are presented by the end of this chapter.

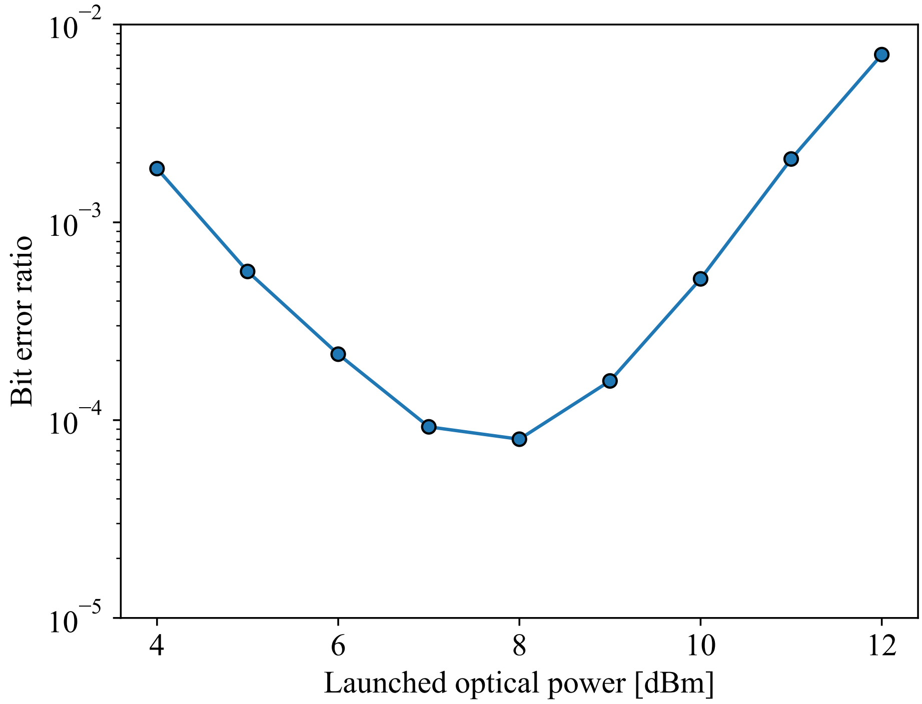
First, a function was implemented for reading the different fiber length possibilities for our dataset. The data is the input of the function which outputs a dictionary with the power level and the two polarization 16-QAM signal values. Then, another function calculates the maximum likelihood result for this signal, as illustrated in Figure 14. Let us first take a closer look at this result.

3.2 Simulation results

For low LOP, the BER is very high. This is mainly because of the nature of the noise associated with the signal. At low power input the stochastic noise is impacting the Signal to Noise Ratio (SNR), and after increasing the optical power we gradually increase the SNR. Nevertheless, after around 10 dBm the BER again starts to increase gradually. This is because now we have the contribution of nonlinearities, and, at that point increasing the LOP no longer does any beneficial effect to our BER. We used a fiber length of 135 kilometers. By the way: the simulation dataset is reutilized from previous works focused on nonlinear distortion mitigation, and it is not our intent to go into the details of this simulation scenario, for it was already explored in previous works (COSTA, 2022).

As a second step, the neural network was trained using 40 repetitions and number of neurons in the hidden layer ranging from 10 to 250. The training and test split was set 50$\%$.

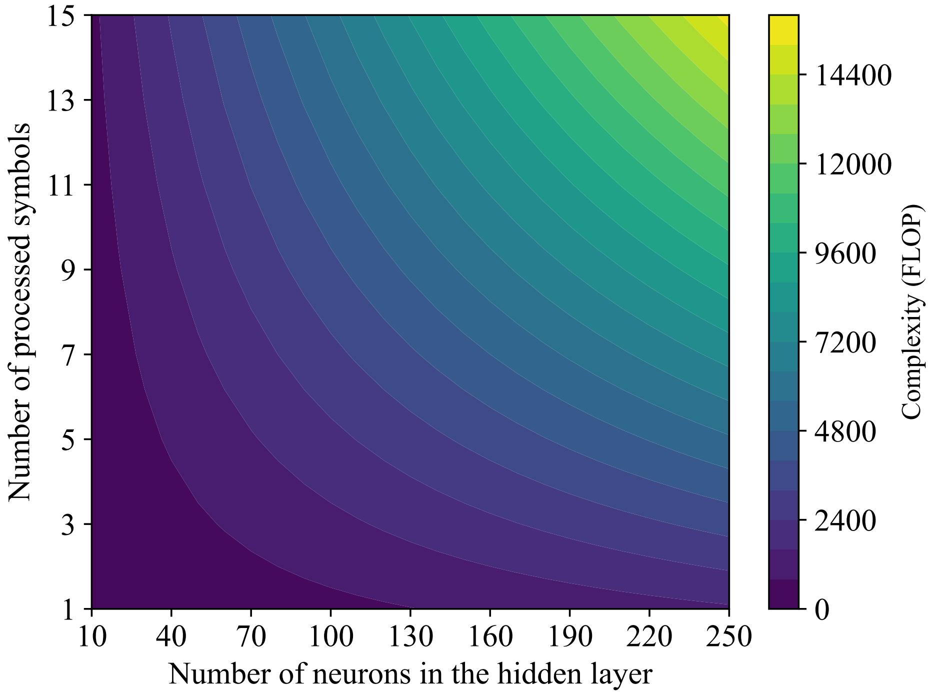
Figure 14 – Bit error ratio results using maximum likelihood method



Source: Author

Then, we analyzed the heatmap comparison, between the BER result, the number of neurons in the hidden layers and number of processed symbols. In the final script, the user can input values for number of symbols, number of repetitions and number of neurons in the hidden layer. The complexity is also calculated, comparing the BER with the different computational complexity of the model used.

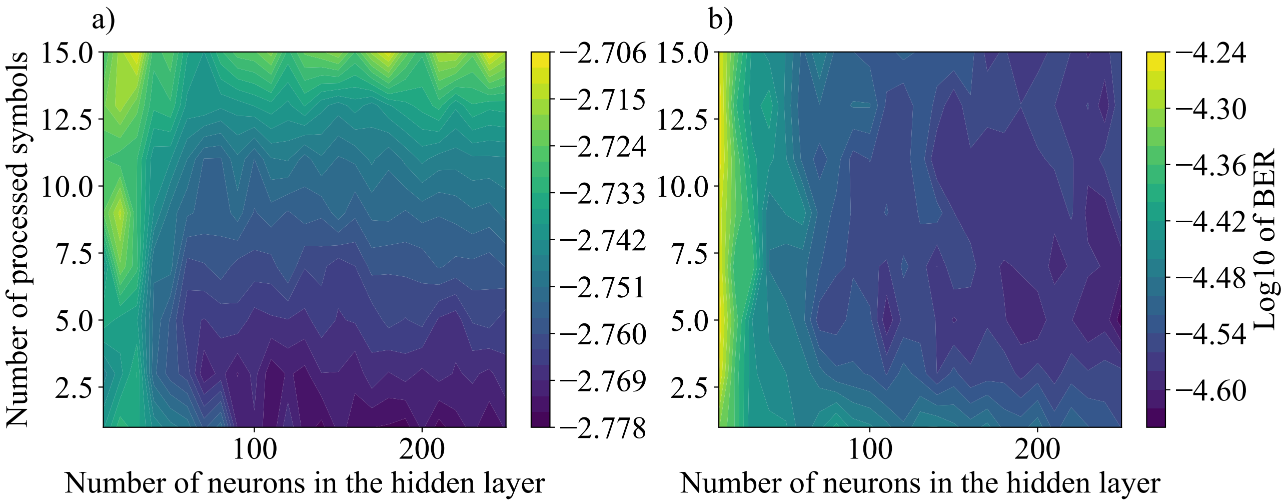
Figure 15 – Computational demand of the neural network



Source: Author

In Figure 15 we can visualize the computational demand of the neural network configuration as both the number of neurons and processed symbols increase. This can help in optimizing the setup by selecting configurations that balance performance and computational load. Next, we can analyze the heatmap for different LOP. Let us start with 4 dBm and 8 dbm in Figure 16.

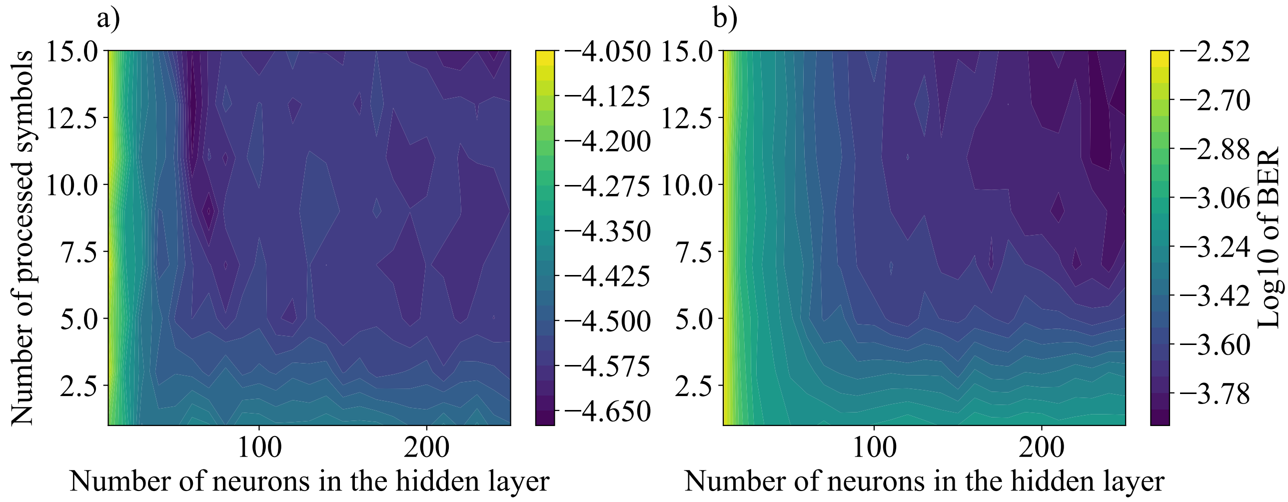
Figure 16 – Bit error ratio variation versus number of processed symbols considering a) 4 dBm input and b) 8 dbm input



Source: Author

For Figure 16 a) the range of BER values does not change much. This indicates that using NNs is not doing any good in this case. This result is expected, because, again, for low power inputs, we are affected by noise and not by the nonlinearity's we are trying to mitigate (as indicated in Chapter 2, the noise is stochastic, while the nonlinearity is deterministic). In other words: our model is not able to mitigate the distortion caused by the telecommunication channel. And that makes sense. Because we are affected mostly by noise, and not, by nonlinearity. When we increase the power input to 8 dbm in Figure 16 b) we already start to observe differences in the range of BER values. We expect that these differences will become more and more pronounced as we increase the LOP. Let us increase the power input to 9 and 12 dBm.

Figure 17 – Bit error ratio variation versus number of processed symbols considering a) 9 dBm input and b) 12 dBm input



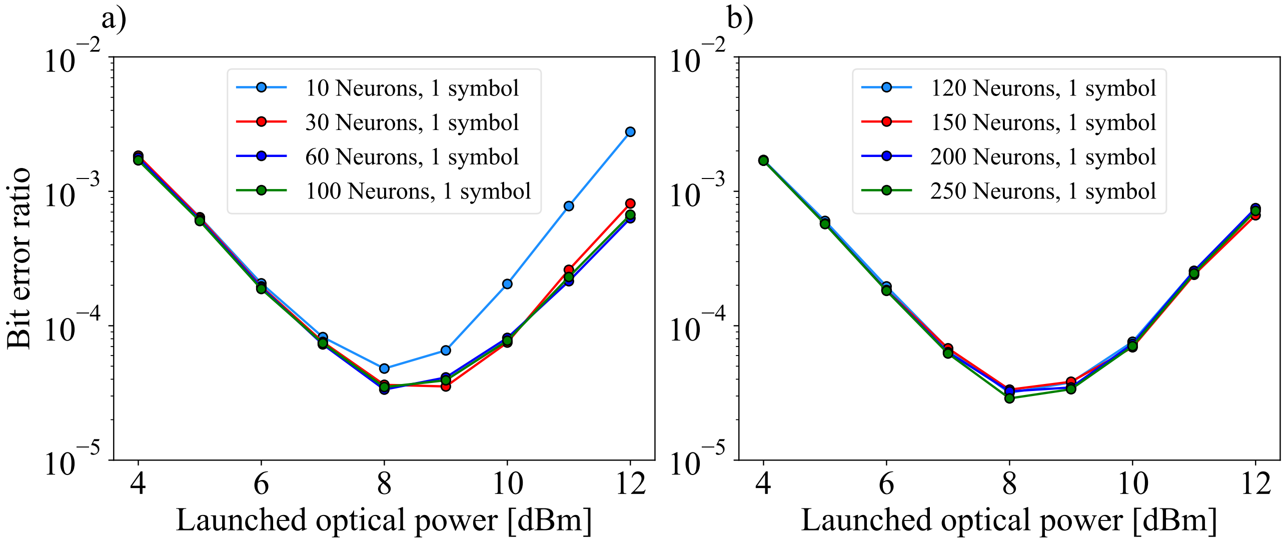
Source: Author

The range of BER values in the heatmap is now increased, as we start to have contributing components of the nonlinear effect. Knowing that, if we increase the input power even more, then we expect our model to have even higher mitigation performance (that is, higher BER range in the heatmap). In the first result, our BER ranges from about 4.24 to 4.60. Now, our range goes from 4.05 to 4.65.

For 12 dBm the BER range is even higher. One important thing to notice is that in Figure 17 a), for number of neurons higher than about 50 the BER starts to get much higher. That can be explained by the model getting *overfitted* (see our remark in Section 2.3).

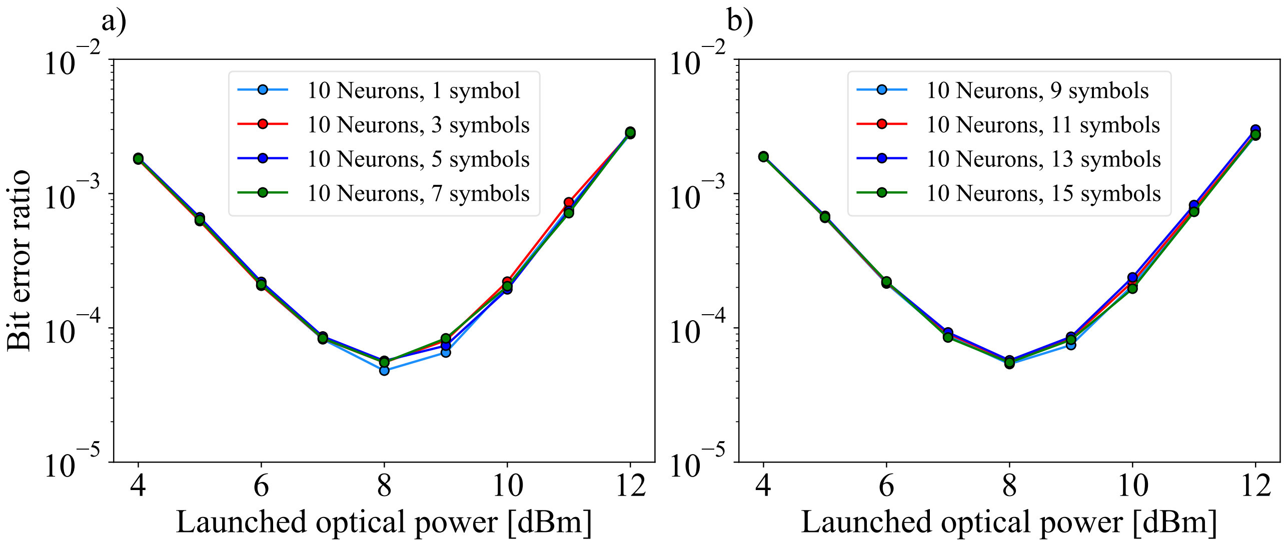
It is also important to visualize the influence of the different number of neurons and number of processed symbols keeping one of the hyperparameters fixed. For that, we also analyzed the differences in the BER that results from varying the number of neurons while keeping the number of processed symbols fixed at 1. And, also, by keeping the number of neurons fixed at 10 and varying the number of processed symbols (we choose 1 and 10 because both would represent the "worst-case" scenario). Figure 19 a) and b) and Figure 18 a) and b) illustrate the results for these two different cases. It is also important to mention here that these illustrations do not contemplate all possible combinations of hyperparameters. The hyperparameters are related to each other, and to understand which regions are optimal for each hyperparameter combination, we recommend Figure 16 combined with Figure 17.

Figure 18 – Bit error ratio variation when changing the number of neurons in the hidden layer considering 1 symbol



Source: Author

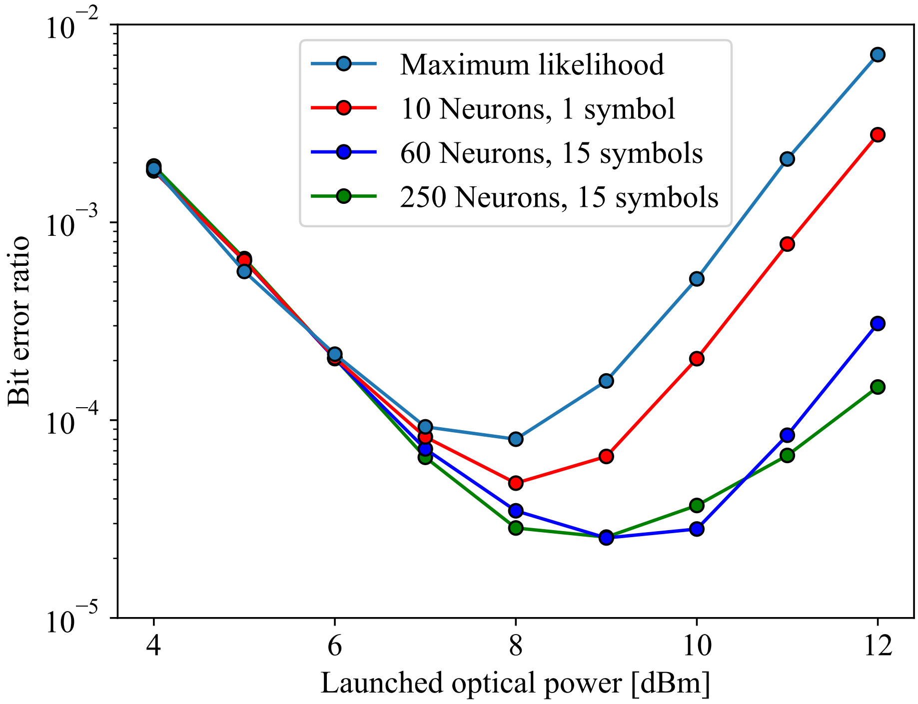
Figure 19 – Bit error ratio variation when changing the number of processed symbols considering 10 neurons for the hidden layer



Source: Author

Let us now compare the BER results for three different cases: 10 neurons and 1 symbol (least effective case), 250 neurons and 15 symbols (highest performance but highest computational cost), and, 15 processed symbols with 60 neurons in the hidden layer. These are good configurations to illustrate our results, because they show the different BER results that are produced by using more processed symbols and also show the optimal result for a LOP of 9 dBm. These configurations are illustrated in Figure 20, also comparing with the results from the maximum likelihood method, previously illustrated in Figure 14.

Figure 20 – Bit error ratio variation considering different number of neurons, symbols and LOP

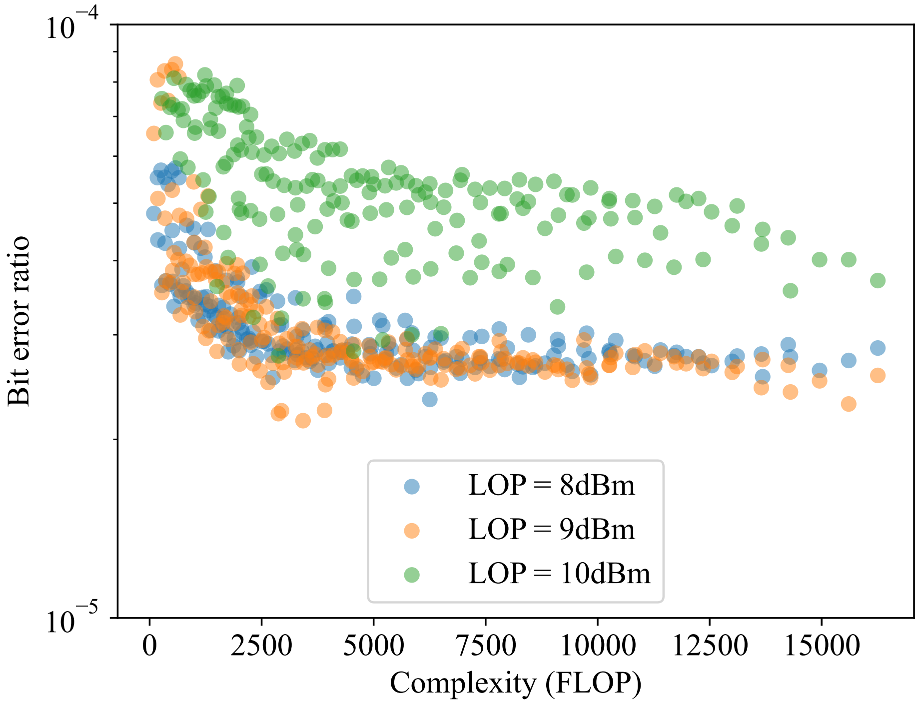


Source: Author

3.3 Pareto front considering bit error ratio and number of operations

One question we may ask is at which computational cost we want to reduce the BER. To address this question, it is also important for us to compare the optimal LOP for each scenario. Figure 21 illustrates the complexity versus the BER for different input powers.

Figure 21 – Training complexity versus bit error ratio comparing LOP



Source: Author

Now, different input power levels result in different BER versus complexity graphical results. Naturally, higher LOP results in more difficult to compensate nonlinear effects, and the lower BER that is targeted, the higher the complexity is needed for the compensation. This graphical visualization allows the user to decide, at which computational cost is the equalizer feasible for reducing the BER to the target value for each power input. Furthermore, the number of operations is represented by the following equations.

The number of neurons in the input layer is given by:

$$N_{input} = 4.N, \tag{3.3}$$

also, the number of neurons in the hidden layer is $N_{neurons}$ and the number of neurons in the output layer is $N_{output} = 4$.

The number of operations in the input-hidden connection is given by:

$$N_{op1} = N_{input} \cdot N_{neurons} + N_{neurons}. \tag{3.4}$$

The number of operations in the hidden-output is given by:

$$N_{op2} = N_{neurons} \cdot N_{output} + N_{output}. \tag{3.5}$$

The total number of operations is given by:

$$N_{op} = N_{op1} + N_{op2} = (N_{input} + 1) \cdot N_{neurons} + (N_{neurons} + 1) \cdot N_{output}, \tag{3.6}$$

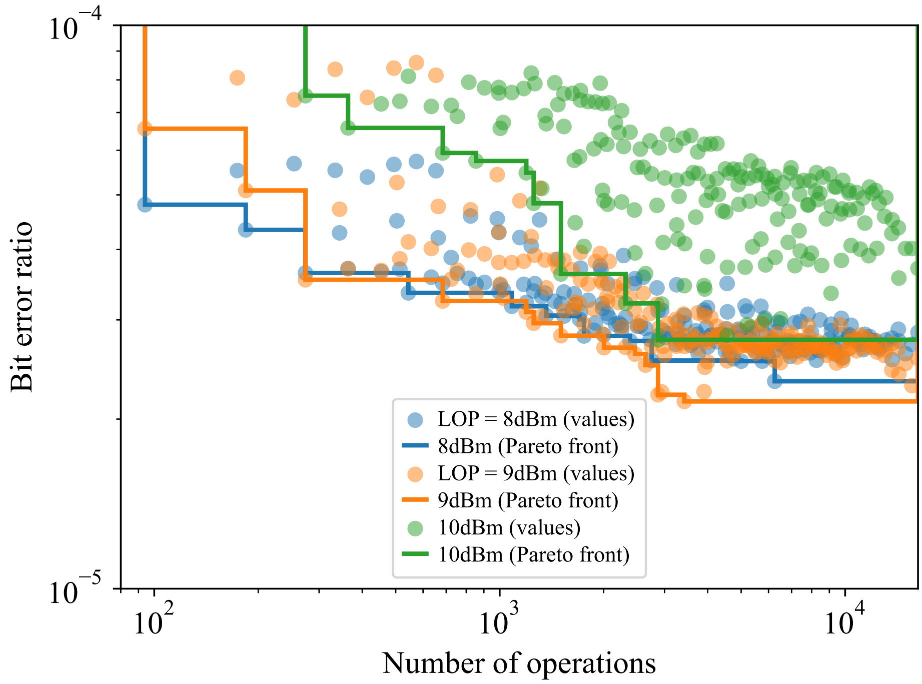
and,

$$N_{op} = (4 \cdot N + 1) \cdot N_{neurons} + 4(N_{neurons} + 1). \tag{3.7}$$

Complexity in units of Floating-Point Operations (FLOP) refers to the number of floating-point calculations needed to complete a task. This unit is commonly used to measure computational complexity. Furthermore, if an algorithm has a complexity of 10^6 FLOPs, it means it performs around 1 million floating-point operations.

The Pareto front is also illustrated in Figure 22, comparing the different dBm values, BER and number of operations needed to perform the training task.

Figure 22 – Pareto front considering bit error ratio versus the number of floating-point operations comparing different LOP



Source: Author

Here, we can observe different behavior for the three cases. The highest input power value requires a much higher number of operations for lowest BER values to be achieved, when compared to the 8 dBm input. This visual representation is very useful for the user, as analyzing this figure, he is able to tell for which input power and at which computational cost it will be possible to mitigate the nonlinear effect and, then, achieve lower desired BER result.

4 CONCLUSIONS AND FUTURE WORK

In this work, we propose the development of a nonlinear equalizer using ANNs. We propose an unusual method, which uses adjacent symbols for training the neural network instead of simply using one symbol, because there is a correlation between consecutive pulses. The dataset used for this nonlinear compensation was simulated using VPI Transmission Maker. We trained the neural network using the in-phase and quadrature constellation values for the 16-QAM signal and compared the results of using the MLP versus the maximum likelihood method. One of the goals of this work is to also take into account the computational complexity of the training task. This is important, because we are using more symbols for the training of the neural network, and this will naturally increase the computational complexity needed for this task. So, two things are considered: the BER comparison between the maximum likelihood method and the ANN, and, the computational complexity needed using different number of neurons in the hidden layer. Our results demonstrate a BER reduction from 10^{-3} to 10^{-4} for 10^2 operations, and that BER can be drastically reduced increasing the number of operations.

As next steps, we consider using RNNs and CNNs, exploring the correlation of adjacent symbols in a broader range of options when compared to using the MLP. With that, we expect to reduce the computational complexity and BER, while considering different fiber lengths and different modulation schemes.

REFERENCES

AGRAWAL, G. P. Nonlinear fiber optics. In: **Nonlinear Science at the Dawn of the 21st Century**. [S.l.]: Springer, 2000. p. 195–211.

AHMED, R. **Machine Learning Regression Masterclass in Python**. 2024. Udemy Video Course. Available at: <https://www.udemy.com/course/machine-learning-regression-masterclass-in-python/>.

AMARI, A. *et al.* A survey on fiber nonlinearity compensation for 400 gb/s and beyond optical communication systems. **IEEE Communications Surveys & Tutorials**, IEEE, v. 19, n. 4, p. 3097–3113, 2017.

AMARI, A. *et al.* A machine learning-based detection technique for optical fiber nonlinearity mitigation. **IEEE Photonics Technology Letters**, IEEE, v. 31, n. 8, p. 627–630, 2019.

ANTIL, R.; PINKI, S. B.; BENIWAL, S. An overview of dwdm technology & network. **Int. J. Sci. Technol. Res**, v. 1, n. 11, p. 43–46, 2012.

AZODOLMOLKY, S. *et al.* Experimental demonstration of an impairment aware network planning and operation tool for transparent/translucent optical networks. **Journal of Lightwave Technology**, IEEE, v. 29, n. 4, p. 439–448, 2011.

BREED, G. Bit error rate: Fundamental concepts and measurement issues. **High Frequency Electronics**, Summit Technical Media, v. 2, n. 1, p. 46–47, 2003.

ÇETIN, O.; TEMURTAŞ, F.; GÜLGÖNÜL, Ş. An application of multilayer neural network on hepatitis disease diagnosis using approximations of sigmoid activation function. **Dicle Tıp Dergisi**, Dicle University, v. 42, n. 2, p. 150–157, 2015.

CHAKRAVORTY, P. What is a signal?[lecture notes]. **IEEE Signal Processing Magazine**, IEEE, v. 35, n. 5, p. 175–177, 2018.

CHO, J. *et al.* How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? **arXiv preprint arXiv:1511.06348**, 2015.

COOLEY, J. W.; TUKEY, J. W. An algorithm for the machine calculation of complex fourier series. **Mathematics of computation**, JSTOR, v. 19, n. 90, p. 297–301, 1965.

COSTA, C. *et al.* Self-phase modulation and inter-polarization cross-phase modulation mitigation in single-channel dp-16qam coherent pon employing 4d clustering. **Optical Fiber Technology**, Elsevier, v. 75, p. 103186, 2023.

COSTA, C. B. Uso de algoritmo de clusterização k-means 4d para mitigação de impedimentos não lineares em sistemas ópticos coerentes digitais dp-16qam. Universidade Estadual Paulista (Unesp), 2022.

DHANASEKAR[1], S.; RAMESH, J. Fpga implementation of variable bit rate 16 qam transceiver system. **International Journal of Applied Engineering Research**, v. 10, n. 10, p. 26497–26507, 2015.

DONGARE, A. *et al.* Introduction to artificial neural network. **International Journal of Engineering and Innovative Technology (IJEIT)**, Citeseer, v. 2, n. 1, p. 189–194, 2012.

DUA, M. *et al.* An improved rnn-lstm based novel approach for sheet music generation. **Procedia Computer Science**, v. 171, p. 465–474, 2020. ISSN 1877-0509. Third International Conference on Computing and Network Communications (CoCoNet'19) Available at: <https://www.sciencedirect.com/science/article/pii/S1877050920310152>.

DURAI, C. R. B. **Digital Signal Processing**. [S.l.]: Laxmi Publications, 2005.

EURONEWS. **'Profound risk to humanity': Elon Musk and Steve Wozniak join calls to halt AI development**. 2023. Disponível em: <https://www.euronews.com/next/2023/03/29/profound-risk-to-humanity-elon-musk-and-steve-wozniak-join-calls-to-halt-ai-development>.

FELDMANN, A. *et al.* The lockdown effect: Implications of the covid-19 pandemic on internet traffic. In: **Proceedings of the ACM internet measurement conference**. [S.l.: s.n.], 2020. p. 1–18.

GAO, Y. *et al.* Intra-channel nonlinearities mitigation in pseudo-linear coherent qpsk transmission systems via nonlinear electrical equalizer. **Optics Communications**, Elsevier, v. 282, n. 12, p. 2421–2425, 2009.

GHOLAMY, A.; KREINOVICH, V.; KOSHELEVA, O. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. **Int. J. Intell. Technol. Appl. Stat**, v. 11, n. 2, p. 105–111, 2018.

GUBRAN, H. Application of artificial neural network for condition monitoring and diagnosis of rotating machinery,. **Hadhramout University Journal of Natural and Applied Sciences**, 01 2013.

HACKELING, G. **Mastering Machine Learning with scikit-learn**. [S.l.]: Packt Publishing Ltd, 2017.

HAYKIN, S. **Communication systems**. [S.l.]: John Wiley & Sons, 2008.

HEBB, D. O. **The organization of behavior: A neuropsychological theory**. [S.l.]: Psychology press, 2005.

JAAFARI, I. E.; ELLAHYANI, A.; CHARFI, S. Rectified non-linear unit for convolution neural network. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2021. v. 1743, n. 1, p. 012014.

JAN, Y.-N.; JAN, L. Y. Dendrites. **Genes & development**, Cold Spring Harbor Lab, v. 15, n. 20, p. 2627–2641, 2001.

KAHN, J. M.; HO, K.-P. Spectral efficiency limits and modulation/detection techniques for dwdm systems. **IEEE Journal of selected topics in quantum electronics**, IEEE, v. 10, n. 2, p. 259–272, 2004.

KAO, K. C.; HOCKHAM, G. A. Dielectric-fibre surface waveguides for optical frequencies. In: IET. **Proceedings of the Institution of Electrical Engineers**. [S.l.], 1966. v. 113, n. 7, p. 1151–1158.

KARMAKER, B. K. Effects of chromatic dispersion and self-phase modulation on optical transmission system. Institute of Information and Communication Technology, BUET, 2009.

KASPAROV, G. **Deep thinking: where machine intelligence ends and human creativity begins**. [S.l.]: Hachette UK, 2017.

LATHI, B. P.; GREEN, R. A. **Linear systems and signals**. [S.l.]: Oxford University Press New York, 2005. v. 2.

LI, M.-J.; LI, S.; NOLAN, D. A. Nonlinear fibers for signal processing using optical kerr effects. **Journal of lightwave technology**, IEEE, v. 23, n. 11, p. 3606, 2005.

LIBRETEXTS. **Raven Biology**. [S.l.]: LibreTexts, 2021.

LIN, G.; SHEN, W. Research on convolutional neural network based on improved relu piecewise activation function. **Procedia computer science**, Elsevier, v. 131, p. 977–984, 2018.

LIN, H. W.; TEGMARK, M.; ROLNICK, D. Why does deep and cheap learning work so well? **Journal of Statistical Physics**, Springer, v. 168, p. 1223–1247, 2017.

LIN, J.-W. Artificial neural network related to biological neuron network: a review. **Advanced Studies in Medical Sciences**, v. 5, n. 1, p. 55–62, 2017.

LUO, X. *et al.* Nonlinear impairment compensation using transfer learning-assisted convolutional bidirectional long short-term memory neural network for coherent optical communication systems. In: MDPI. **Photonics**. [S.l.], 2022. v. 9, n. 12, p. 919.

MARGRIE, T. W.; URBAN, N. Dendrites as transmitters. **Dendrites**, Oxford University Press, USA, p. 401, 2007.

MATA, J. *et al.* Artificial intelligence (ai) methods in optical networks: A comprehensive survey. **Optical switching and networking**, Elsevier, v. 28, p. 43–57, 2018.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, p. 115–133, 1943.

MEIßL, F. *et al.* Online handwriting recognition using lstm on microcontroller and imu sensors. In: **2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)**. [S.l.: s.n.], 2022. p. 999–1004.

MELLO, D. A. de A.; BARBOSA, F. A. **Digital Coherent Optical Systems**. [S.l.]: Springer, 2021.

MINSKY, M. The age of intelligent machines: thoughts about artificial intelligence. **KurzweilAl. net**, KurzweilAl. net, 1990.

MINSKY, M.; PAPERT, S. An introduction to computational geometry. **Cambridge tiass., HIT**, v. 479, n. 480, p. 104, 1969.

MORAIS, R. M. *et al.* Genetic algorithm for the topological design of survivable optical transport networks. **Journal of optical communications and networking**, Optica Publishing Group, v. 3, n. 1, p. 17–26, 2011.

MUKHERJEE, B. Wdm optical communication networks: progress and challenges. **IEEE Journal on Selected Areas in communications**, IEEE, v. 18, n. 10, p. 1810–1824, 2000.

OKOSHI, T. Recent advances in coherent optical fiber communication systems. **Journal of lightwave technology**, IEEE, v. 5, n. 1, p. 44–52, 1987.

OPPENHEIM, A. V. **Discrete-time signal processing**. [S.l.]: Pearson Education India, 1999.

ORUH, J.; VIRIRI, S.; ADEGUN, A. Long short-term memory recurrent neural network for automatic speech recognition. **IEEE Access**, v. 10, p. 30069–30079, 03 2022.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PROAKIS, J. G.; SALEHI, M. **Digital communications**. [S.l.]: McGraw-hill, 2008.

RAFIQUE, D. *et al.* Compensation of intra-channel nonlinear fibre impairments using simplified digital back-propagation algorithm. **Optics express**, Optica Publishing Group, v. 19, n. 10, p. 9453–9460, 2011.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

ROUDAS, I. Coherent optical communication systems. In: **WDM Systems and Networks: Modeling, Simulation, Design and Engineering**. [S.l.]: Springer, 2011. p. 373–417.

RUMELHART, D. E. *et al.* **Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations**. [S.l.]: The MIT press, 1986.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Pearson, 2016.

SANDLER, U.; TSITOLOVSKY, L. **Neural cell behavior and fuzzy logic**. [S.l.]: Springer, 2008.

SAVORY, S. J. Digital filters for coherent optical receivers. **Optics Express**, Optica Publishing Group, v. 16, n. 2, p. 804–817, 2008.

SCHAEFFER, J.; LAKE, R. Solving the game of checkers. **Games of no chance**, Cambridge University Press Cambridge, v. 29, p. 119–133, 1996.

SCHMIDGALL, S. *et al.* Brain-inspired learning in artificial neural networks: a review. **arXiv preprint arXiv:2305.11252**, 2023.

SCHOFIELD, H. How napoleon's semaphore telegraph changed the world. **BBC, June**, 2013.

SENIOR, J. M.; JAMRO, M. Y. **Optical fiber communications: principles and practice**. [S.l.]: Pearson Education, 2009.

SHRADDHA, N. B.; VIKAS, U. D.; SHANTANU, S. J. Analysis of spm, xpm, and fwm in fiber optic comunication using optisystem. **IJSTE-International Journal of Science Technology & Engineering**, v. 2, n. 07, 2016.

SILVA, I. N. d.; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais para engenharia e ciências aplicadas. 2010.

SNITZER, E. Cylindrical dielectric waveguide modes. **JOSA**, Optica Publishing Group, v. 51, n. 5, p. 491–498, 1961.

SOMAN, S. K. O. A tutorial on fiber kerr nonlinearity effect and its compensation in optical communication systems. **Journal of Optics**, IOP Publishing, v. 23, n. 12, p. 123502, 2021.

STEIN, M. L. **Interpolation of spatial data: some theory for kriging**. [S.l.]: Springer Science & Business Media, 2012.

TURING, A. M. Intelligent machinery, a heretical theory. **Philosophia Mathematica**, OUP, v. 4, n. 3, p. 256–260, 1996.

WERTS, A. Propagation of coherent light in optical fibers(electromagnetic theory of wave propagation modes in dielectric cylinders applied to coherent light transmission in optical fibers). **L'onde Electrique**, v. 46, p. 967–980, 1966.

WINZER, P. J. Optical networking beyond wdm. **IEEE Photonics Journal**, IEEE, v. 4, n. 2, p. 647–651, 2012.

XIAO, Z. *et al.* A multi-class objects detection coprocessor with dual feature space and weighted softmax. **IEEE Transactions on Circuits and Systems II: Express Briefs**, IEEE, v. 67, n. 9, p. 1629–1633, 2020.

ZHANG, S. *et al.* Field and lab experimental demonstration of nonlinear impairment compensation using neural networks. **Nature communications**, Nature Publishing Group UK London, v. 10, n. 1, p. 3033, 2019.

APPENDIX A — SIGNALS

Signals are associated with the passing of information through a medium, (with the purpose of communication between humans and other humans or machines) and can be written mathematically as functions of one or more independent variables (By the way: a speech signal, is represented as a function of time). Then, the independent variable can be continuous or discrete. “Continuous signals” are what we write as analog signals. “Discrete-time signals” are defined at discrete times and the variable will have discrete values (*i.e.*, represented as a sequence of numbers). “Digital signals”, then, are those composed of discrete amplitude and time values (OPPENHEIM, 1999).

The authors know that signals consists of sets of data, and that includes closing stock prices, voice information coming through a public telephone or from the television (audio, video or speech). In this work, signals are going to be functions of the variable time (LATHI; GREEN, 2005). And this is important too: a signal can be any quantity that varies over space and time to share messages between observers (CHAKRAVORTY, 2018).

Digital Signal Processing (DSP) corresponds to various techniques which use the representation of signals by sequence of numbers and symbols and the processing of the sequence. Most applications of DSP include Bio-medical engineering, radar or nuclear science (DURAI, 2005).

There are five methods of classifying signals based on different features.

- Based on independent variable
  - Continuous time signals: The signal is represented continuously through time, *i.e.*, we say it is continuous if it is defined for all time.
  - Discrete time signals: A discrete time signal is represented by a finite number of digits, *i.e.*, the independent variable will have discrete values only.
- Depending on the number of independent variables
  - 1-D signals: It is function of a single independent variable. One example is voice speech signal.
  - 2-D signals: It is a function of two or more independent variables, e.g.: image signal.
  - M-D signals: It is a function of M independent variables in time, e.g.: video signals.
- Depending on the certainty by which the signal can be described

- Deterministic Signal: A signal that we can write as a mathematical expression or rule. One example we can give is the sinusoidal signal, written as:

$$v(t) = V_m \cdot \sin(2\pi f t) \qquad \text{for} \qquad t \geq 0. \tag{A.1}$$

  A square signal can be represented by:

$$x(t) = A \qquad \text{for} \qquad 0 \leq t \leq T/2, \tag{A.2}$$

$$x(t) = -A \qquad \text{for} \qquad T/2 \leq t \leq T. \tag{A.3}$$

- Random signal: A signal that cannot be predicted, examples are speech signals and electrocardiogram signals.

• Based on repetition

- Periodic Signal: A periodic signal is a signal that is repeated every finite interval of time, or a continuous time function that satisfies the following condition:

$$x(t) = x(t + T). \tag{A.4}$$

  The smallest period of T is defined as the *Fundamental period*. We can also define the *Fundamental frequency* as:

$$f = \frac{1}{T}. \tag{A.5}$$

  Then, for a discrete time signal, it is said to be periodic if it satisfies the following equation:

$$x[n] = x[n + N], \tag{A.6}$$

  where N is a positive integer.

- Non-periodic signals: any signal that does not satisfy Equation A.6.

• Based on reflection.

- Even signals: a signal $x(t)$ or $x[n]$ is an even signal if it is identical to its time reverse, that is:

$$x(t) = x(-t). \tag{A.7}$$

– Odd signals: a signal $x(t)$ or $x[n]$ is an odd signal if:

$$x(t) = -x(-t). \tag{A.8}$$

APPENDIX B — MODULATION AND 16-QAM

The waveform used to transmit data can appear in different forms, and we are going now to describe the Pulse amplitude modulation (PAM). For digital PAM, the signal waveform can be written as Equation B.1 (PROAKIS; SALEHI, 2008).

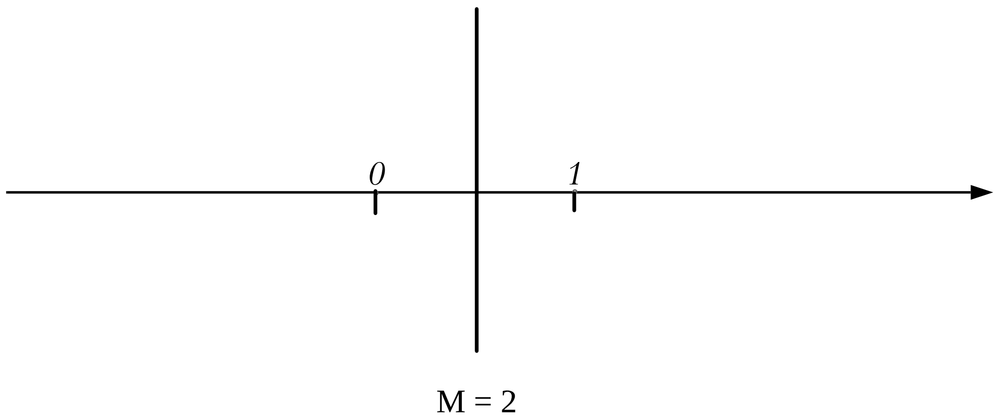
$$s_m(t) = A_m p(t), \qquad 1 \leq m \leq M, \tag{B.1}$$

now, $p(t)$ corresponds to the pulse with duration T and $A_m, 1 \leq m \leq M$, possible amplitudes that corresponds to $M = 2^k$ possible k-bit blocks of symbols. We know that quite often, signal amplitudes assume discrete values, as described in Equation B.2.

$$A_m = 2m - 1 - M, \qquad m = 1, 2, ..., M. \tag{B.2}$$

Then, amplitudes are: $\pm 1, \pm 2, \pm 5, ..., \pm (M-1)$. In this description there is no carrier modulation involved, but in many cases, there is the need to modulate signals. Figure 23 illustrates PAM signalling. Band pass signal PAM also receives the name Amplitude shift keying (ASK).

Figure 23 – Constellation for PAM signal



Source: PROAKIS, 2018

The modulation process is how we are going to move the message signal to one specific frequency band. One important remark now: the "*carrier*" is a sinusoidal signal of high frequency. Using modulation, one of the carrier parameters (which can be: amplitude, frequency or phase) is changed, in proportion to our base band signal. The term "baseband" is the frequency band of the original message signal from our source. Baseband signals have overlapping bands. For that reason, they leave great part of the channel spectrum unused, and this is one of the reasons why it is important to modulate signals. A remark of extreme importance: the assignment of bit combinations to symbols is not arbitrary. One particular coding scheme known as Gray coding will map the bits in a certain way that only a single bit changes between two adjacent symbols in the constellation diagram. If this is not done, one symbol error can cause multiple bit errors resulting in BER increase (AGRAWAL, 2000).

The bandwidth efficiency of PAM can be attained by suppressing two k-bit symbols from the information sequence on two quadrature carriers $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$. The resulting modulation scheme is called Quadrature Amplitude Modulation (QAM). The signal waveform can be written as in Equations B.3 and B.4.

$$s_m(t) = Re[(A_{mi} + jA_{mq})g(t)e^{j2\pi f_c t}] \tag{B.3}$$

$$= A_{mi}g(t)\cos(2\pi f_c t) - A_{mq}g(t)\sin(2\pi f_c t), \qquad m = 1, 2, ..., M. \tag{B.4}$$

Now, A_{mi} and A_{mq} are the information signal amplitudes of quadrature carriers and $g(t)$ corresponds to the signal pulse. Then, we can write:

$$s_m(t) = \mathrm{Re}\left[r_m e^{j\theta_m} e^{j2\pi f_c t}\right] \tag{B.5}$$

$$= r_m \cos\left(2\pi f_c t + \theta_m\right). \tag{B.6}$$

Then,

$$r_m = \sqrt{A_{mi}^2 + A_{mq}^2}, \tag{B.7}$$

and,

$$\theta = \tan^{-1}\left(\frac{A_{mq}}{A_{mi}}\right). \tag{B.8}$$

It then becomes clear that QAM signals are composed of amplitude and phase modulation combination.

The Euclidean distance between a pair of signal vectors in QAM is given by the Equation B.9.

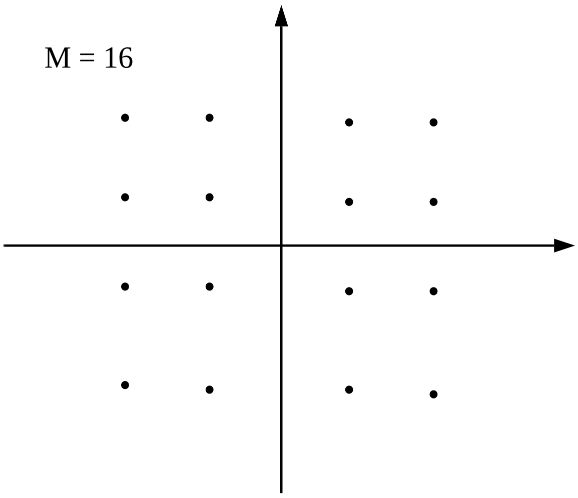
$$d_{mn} = \sqrt{||s_m - s_n||^2} = \sqrt{\frac{\mathcal{E}_g}{2}[(A_{mi} - A_{ni})^2 + (A_{mq} - A_{nq})^2]}, \tag{B.9}$$

where $\mathcal{E}_g$ is the average energy per symbol of the signal constellation. In the case where signal amplitudes assume discrete values the signal space is rectangular. In this case, the Euclidean distance is given by:

$$d_{min} = \sqrt{2\mathcal{E}_g}. \tag{B.10}$$

Finally, when $M = 16$ we have what is called a 16-QAM signal. Figure 24 illustrates this case.

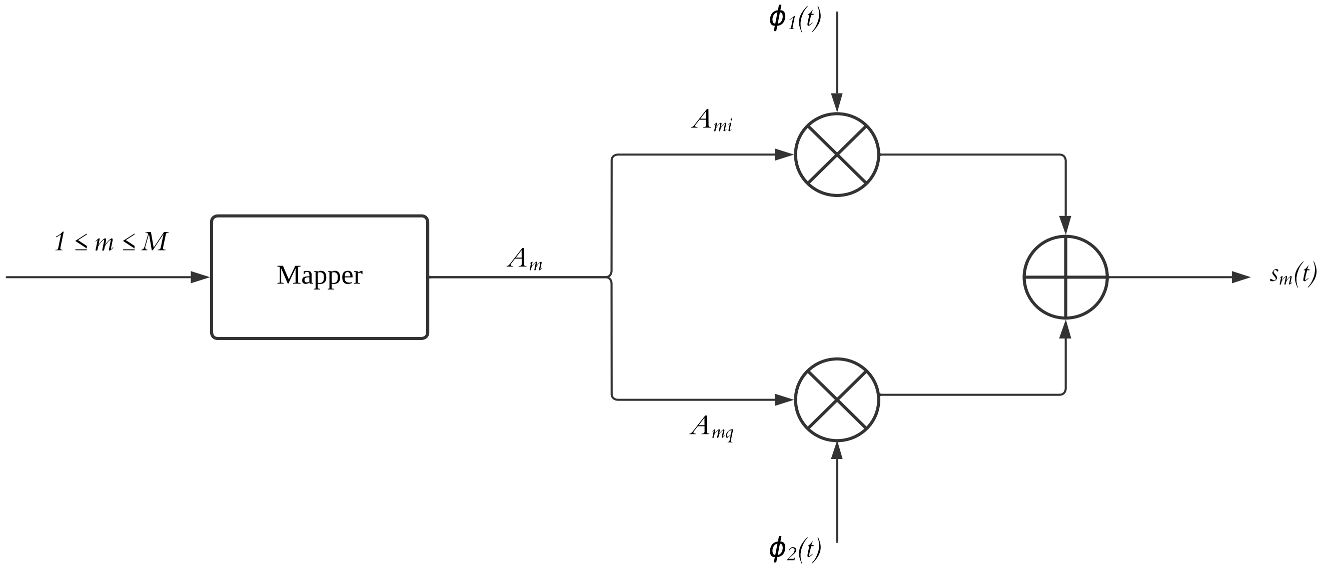
Figure 24 – 16-QAM signal



Source: PROAKIS, 2018

Figure 25 illustrates a quadrature amplitude modulator.

Figure 25 – Quadrature amplitude modulator



Source: PROAKIS, 2018

One more clarification: $\phi_1(t)$ and $\phi_2(t)$ are given by:

$$\phi_1(t) = \sqrt{\frac{2}{\mathcal{E}_g}} g(t) \cos(2\pi f_c t), \tag{B.11}$$

and,

$$\phi_2(t) = \sqrt{\frac{2}{\mathcal{E}_g}} g(t) \sin(2\pi f_c t). \tag{B.12}$$

In the case of QAM a constellation diagram is repeatedly used for analysis (DHANASEKAR[1]; RAMESH, 2015). The constellation diagram refers to a plot of the symbols in two-dimensional space. For 16-QAM signals, we are going to have 16 symbols. Each symbol consists of 2 bits. 16-QAM signals have 4 amplitudes and 12 phases, assuming the four discrete values: 0.5, -0.5, 1.5 and -1.5. The data is divided into two streams: the in-phase (I) and the quadrature (Q).

APPENDIX C — WAVELENGTH-DIVISION MULTIPLEXING

Wavelength-Division Multiplexing (WDM) is an approach that takes into advantage the big opt-electronic bandwidth, by dividing the optical spectrum into a certain number of non-overlapping wavelength bands. This way, each separated channel operates at the desired rate and with relatively easy implementation (MUKHERJEE, 2000). WDM networks were deployed around the early 1990s and by early 2000s lasers have reached Gigahertz frequency stability (WINZER, 2012).

Dense Wavelength-Division Multiplexing (DWDM) is an extension of optical networking. Instead of requiring one optical fiber per transmitter, DWDM allows multiple channels to occupy a single fiber-optic cable. They can carry different types of traffic at different speeds (ANTIL; PINKI; BENIWAL, 2012).

APPENDIX D — BIT ERROR RATIO

For a digital communication system, the channel is used as the physical medium to bring the signal from the transmitter to the receiver. Our used measurement for digital communication systems performance is what we call the Bit Error Ratio (BER). BER can be defined in a very simple form as:

$$\text{BER} = \frac{Errors}{Number \quad of \quad bits}. \tag{D.1}$$

For this to be true, BER is a unitless measurement, which can be affected by transmission channel noise and interference. Noise is contrary to the BER, as more noise results in an increased BER. Additionally, this same noise is a random process, defined in terms of statistics. And this has to do with the fact that it can be described by a Gaussian distribution function. With a strong signal and unperturbed signal path, this number is small and not significant (BREED, 2003).

There is one way to lower spectral noise density by reducing bandwidth, but we are limited by the bandwidth required to transmit the desired bit rate (Nyquist criterion [1]) (OPPENHEIM, 1999). We can also increase the energy per bit by using higher power transmission, but interference with other systems can limit that option. A lower bit rate increases the energy per bit, but then there is a loss in capacity.

As example let us assume the following bit sequence:

```
110001011
```

and the received sequence:

```
010101001
```

In this case, the BER is 3 incorrect bits divided by 9 transmitted bits, and that is equal to a BER of 0.333 or 33.3 % .

[1] To reconstruct a continuous signal, the sampling frequency (rate) must be at least twice the highest frequency component present in the signal, that is: $f_s \geq 2f_{max}$.

APPENDIX E — NOISE

White noise is a random signal. It receives the name “white” because it is present in all frequencies (in reference to visible light, white light is light composed of all frequency components) (STEIN, 2012). An infinite-bandwidth white noise signal is a theoretical construction. The bandwidth of white noise is limited in real world applications, and the noise itself is stochastic and cannot be compensated. It is present in electronic and communication systems.

There are more types of noise present in communication systems, such as shot noise and thermal noise.

Shot noise comes from electronic devices (diodes and transistors). One example is the photodetector. In a photodetector, the current pulse is generated every time an electron is emitted by the cathode due to light incidence. The electrons are emitted at random times. Then, the number of electrons $N(t)$ emitted in a time interval $[0, t]$ is a discrete stochastic process.

Thermal noise comes from the random motion of electrons in a conductor. The mean square value of thermal noise voltage across the terminals of a resistor is given by:

$$E[V_{tn}^2] = 4kTR\Delta f. \tag{E.1}$$

Here, k is Boltzmann’s constant, T is the absolute temperature in Kelvin and R is the resistance in ohms (HAYKIN, 2008). $E[V_{tn}^2]$ values are in $volts^2$.

Phase noise is created during the process of combination and recombination of charge carriers inside the molecular structure of the semiconductor. Instantaneous frequency and phase are related by the formula:

$$f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt}. \tag{E.2}$$